Blind Adaptive Signal Processing with Applications to Channel Equalization and WDM Fiber-Optic Receivers

A THESIS

Presented to the Academic Faculty

By

Michael Joseph Minardi

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering

Georgia Institute of Technology

School of Electrical and Computer Engineering

March 1996

Blind Adaptive Signal Processing with Applications to Channel Equalization and WDM Fiber-Optic Receivers

Approved:

Professor Mary Ann Ingram, Chairman

Professor John Barry

Professor Douglas Williams

Date approved by Chairman Dec 18, 1995

Dedicated with love to my wife

Laurie,

and to my children

Rebecca, Jack, Kellyann, Dianne and Joseph.

ACKNOWLEDGEMENTS

I would not have been able to finish this research without the help of many people. First, I thank my advisor Professor Mary Ann Ingram for her guidance and encouragement. I especially appreciate her understanding during the temporary slowdowns following the births of three of my children which occurred during our collaboration. I would also like to thank the other members of my thesis committee, Professors John Barry, Douglas Williams, Daniel Blumenthal and James Herod. Their comments and suggestions have greatly improved the document you now hold in your hands. I thank my fellow students who's stimulating conversations have enriched my time here at Georgia Tech. Among them are: Paul Anderson, Ann Collier, Tony Wang, Johan Kullstam, Scott Goldstein, Pete Bergstrom, Ho Yang, Steven Hill and Anuj Batra. I especially would like to thank Johan for giving me a copy of his father's unpublished manuscript "The Pseudo-Inverse, the Great Problem Solver."

I had two employers who's support made this thesis possible. The Georgia Tech Research Institute let me work part time and paid for my tuition. I thank my supervisors, Armand Masse, Harry Andrews and Charlie Krebs (who is smiling up in heaven) and my coworkers, Jack Landgren, Fred McKeen, Mark Smith, Alex Howard, Bob Wohlers, Jason Collins, Jerry Heckman, Mike Baden, Marv Cohen and all the great people of the Countermeasures Development Laboratory for their encouragement and support. For the last two years the US Air Force and its Palace Knight program made it possible for me to study full time. I thank my mentor Emil Martinsek, my supervisors Don Cambell and Mike Gardner and Lu Voronyak and Betty Ferguson of the Palace Knight program.

Finally, none of this would have been possible without the support and inspiration of my family. Foremost, my beautiful wife Laurie who would not let me quit on my dream

of a Ph.D., even during the dark days of around-the-clock feeding and diaper changing of the twins. The five apples of my eye, Rebecca, Jack, Kellyann, Dianne, and Joseph, who would occasionally let daddy alone to do his "homework." I want to thank my parents Dr. John Minardi and Joan Ferris Minardi. My mother gave me my faith in God and an optimistic attitude. My father gave me a love of knowledge and showed me what it means to be a man. Dad also gave me some special inspiration, whenever I thought it could not be done with five kids, I would recall my dad graduating with his Ph.D. while his *eight* children looked on.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

$\mathbf{a}$	Vector of data symbols
a_m	Data symbol sequence
AR	Auto-Regressive
ARMA	Auto-Regressive Moving Average
AWGN	Additive White Gaussian Noise
$\mathbf{B}$	Signal blocking matrix
BER	Bit Error Rate
BPSK	Binary Phase Shift Keying
$\mathbf{C}$	Constraint vector
CDD	Constrained Decision Directed Algorithm
CMA	Constant Modulus Algorithm
DDA	Decision Directed Algorithm
$E\{\bullet\}$	Expectation
e_m	Bussgang pseudo-error, y_m-$g(y_m)$
$\mathbf{e}(\theta)$	RF array steering vector for a plane wave from direction θ
FIR	Finite Impulse Response
$g(\bullet)$	Bussgang nonlinearity
$\mathbf{G}$	Crosstalk gain matrix
GSC	Generalized Sidelobe Canceller
$\mathbf{H}$	Channel matrix
h_m	Channel Impulse Response
HOS	Higher Order Statistic
i.i.d.	Independent identically distributed
IIR	Infinite Impulse Response
$\mathbf{I}_d$	d^{th} column of the identity matrix, all zeros except for a one for the d^{th} position

ISI	Inter Symbol Interference
J	Cost function
LMS	Least Mean Squares
LTI	Linear Time Invariant
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
MUSIC	MUltiple SIgnal Classification algorithm
$\mathbf{n}$	Vector of noise
n_c	Number of WDM channels
n_d	Number of WDM crosstalk cancellation weights
OOK	On off Keying
$\mathbf{P}_\perp$	Projection operator onto the nullspace of $\mathbf{H}$
PAM	Pulse Amplitude Modulation
PSD	Power Spectral Density
QAM	Quadrature Amplitude Modulation
$\mathbf{R}$	Autocorrelation matrix
RF	Radio Frequency
RLS	Recursive Least Squares
$\mathbf{S}$	Vector of WDM channel signals
SCD	Spectral Correlation Density
SCNR	Signal to Crosstalk plus Noise Ratio
SINR	Signal to Interference plus Noise Ratio
SIR	Signal to Interference Ratio
SIR_d	Maximum SIR for a delay of d samples
$\mathbf{T}$	Vector of overall receiver/canceller or channel/weight response
t_m	Overall channel equalizer impulse response
ULA	Uniform Linear Array
VC	Variance-Constraint algorithm
$\mathbf{W}$	Weight vector for canceller or equalizer
$\mathbf{W}_{MSE}$	Weight vector that produces minimum MSE

$\mathbf{W}_{SCNR}$	Weight vector that produces maximum SCNR
$\mathbf{W}_{dd}$	Weight vector that is a stable solution for CDD algorithm
WDM	Wavelength Division Multiplexing
x_m	Channel output
y	Equalizer or canceller output
$\mathbf{Z}$	Vector of WDM receiver detector currents
$\nabla_{\mathbf{W}}$	Gradient with respect to $\mathbf{W}$
μ	Stochastic algorithm step size, a positive real number
σ^2	Noise power
θ	Plane wave angle of arrival
$(\bullet)_d$	d^h column of a matrix
$(\bullet)^H$	Hermetian transpose of a matrix
$(\bullet)^T$	Transpose of a matrix
$(\bullet)^{\#}$	Penrose pseudo-inverse of a matrix

SUMMARY

We apply blind adaptive signal processing to two applications, equalization of communications channels and crosstalk cancellation in wavelength-division multiplexed (WDM) fiber optic receivers.

An important limiting factor in WDM fiber-optic network design is crosstalk between channels. Many types of demultiplexers make an array of channel signals available to the receiver. We propose to eliminate the crosstalk by processing these electrical signals after detection. We developed a suite of adaptive algorithms. Some of the concepts used are borrowed from two mature technologies, adaptive array processing and blind channel equalization. Simulations show that this simple signal processing technique in the receiver can more than double the usable capacity of a WDM optical link. For example, a simulation increased an initial 2.5 dB signal-to-crosstalk-ratio to over 35 dB. Some of the algorithms automatically adapt to lasers that drift in frequency, another problem in WDM networks.

Current Bussgang-type blind equalization algorithms have well-known problems with misconvergence. Misconvergence is defined as stable weights that have high residual inter-symbol interference. We show that the weights that cause the overall channel-equalizer response to approximate a pure delay (in a mean square error sense) are within the region of convergence of a stable local minimum of the Bussgang cost function. For any nontrivial channel the approximations for some delay values are poor. If a Bussgang algorithm is initialized with these weights it misconverges. We performed a Monte-Carlo simulation with over 3000 channels of two popular Bussgang algorithms, Decision Directed and Constant Modulus. In all cases we created misconvergence as predicted.

We developed the Variance Constraint (VC) algorithms, a new class of blind equalizers that have more robust convergence properties than the Bussgang algorithms. The VC cost functions are similar to the well-known Godard cost functions, but they incorporate estimates of the channel-equalizer overall gain. They also use a linear constraint to restrict the weights to an affine space. These two modifications push the bad stable points farther away in weight space from the stable points with good performance (i.e., low inter-symbol interference). A Monte-Carlo simulation of nearly 2000 channels shows not a single case of misconvergence.

CHAPTER 1.

INTRODUCTION

The term "blind", when used in reference to adaptive signal processing, means adaptation without training. In this thesis we apply blind adaptive signal processing to two problems: channel equalization in digital receivers and crosstalk cancellation in wavelength division multiplexed (WDM) fiber-optic receivers. Blind equalizer algorithms were first proposed in [Sato '75] and are used in modems and digital receivers. Blind algorithms for WDM crosstalk cancellation were first proposed in [Minardi '92].

Inter-symbol interference (ISI) arises in digital communications when the basic digital waveform is dispersed in time as it propagates through the channel. ISI may occur because of bandwidth limitations of the channel or from other effects like multipath dispersion during signal propagation in wireless communications. Equalization is the process of filtering the received signal to remove ISI. Equalization is required to achieve reliable performance in many high-speed digital communications applications. Blind adaptive algorithms determine the proper filter solely by examining the output of the channel and equalizer. Blindness is desirable because it does not require any cooperation between the transmitter and receiver.

Figure 1 shows a block diagram of a channel and adaptive equalizer. A transmitter sends a sequence of symbols through a linear time invariant channel. The channel output plus noise then passes through a finite impulse response (FIR) filter with adaptive weights. We desire that the weights create an overall channel-equalizer impulse response equal to a pure delay.

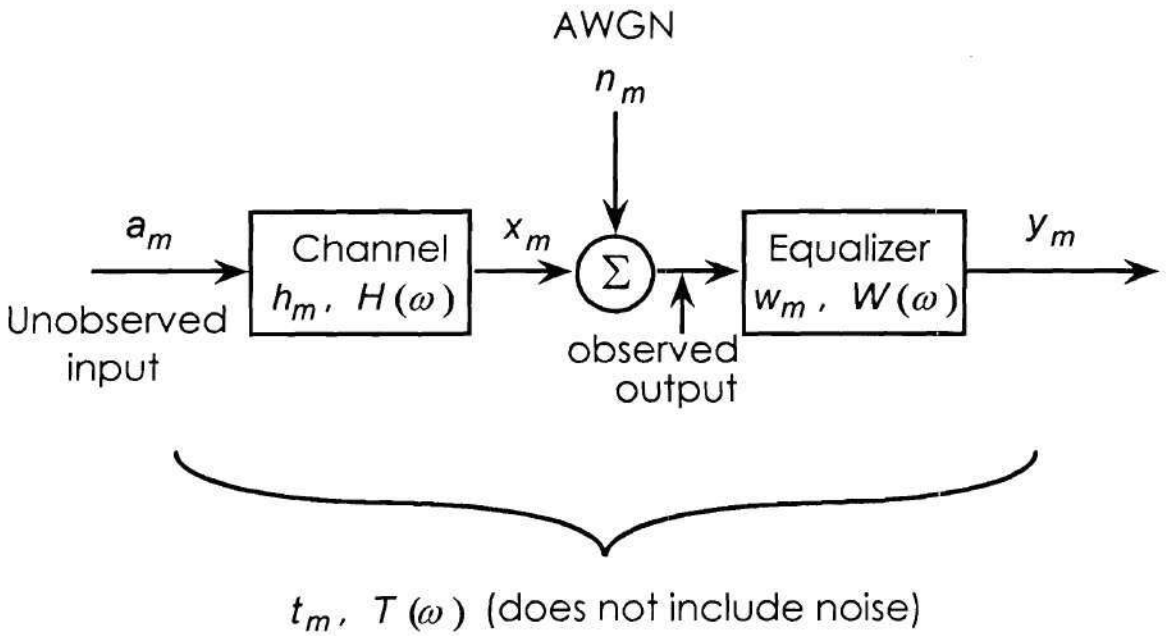


Figure 1: Block diagram of channel with equalizer.

The huge bandwidth capacity of single mode fibers (> 50 THz) suggests that a single fiber may serve many users. Traditional solutions have been to divide the users into time slots, for time-division multiplexing (TDM), wavelength slots for wavelength-division multiplexing (WDM), or to assign each user a unique code for code-division multiplexing (CDM).

TDM requires that all the light pulses are converted to electrical impulses before they are demultiplexed. Therefore, the total data rate of all channels cannot exceed the maximum speed of electronic switches (currently about 10 GHz). This problem is known as the "electronic bottleneck". TDM also requires that all data sources must be synchronized in time.

Many CDM schemes have been proposed, but they all contain major weaknesses. So-called optical orthogonal codes [Chung '89, Salehi '89, Pruncal '86] are extremely

sparse (duty cycles of at most a few percent) and have to be enormously long (code lengths greater than 1000 chips) to accommodate even a few tens of users. Sequence Inversion Keying [Tamura '85, O'Farrell '89] permits noncoherent systems to use the standard codes from classical spread spectrum. These codes are not so sparse and usually have about 50% ones. Chip size is limited to about a nanosecond because electronic components are an inherent part of the matched filter. In other words, they suffer from the electronic bottleneck.

For these and other reasons, interest in CDM for fiber-optic networks is waning. For example, the most recent Optical Fiber Communications conference (OFC '95) did not contain a single paper on the subject of CDM, but contained six sessions on WDM networks. Apparently, the leading candidate for the "information highway" of the future is WDM.

One of the major problems for WDM networks is crosstalk between the channels [Hill '85]. At infrared wavelengths, good demultiplexers require channel separations of 1 nm to achieve the 20 dB signal-to-crosstalk-plus-noise ratio (SCNR) required for a 10^{-9} bit error rate (BER) [Geckeler '87]. One nm represents a required channel separation of about 140 GHz at λ=1.5 μm. Assuming a maximum channel bandwidth of 10 Ghz, only 7 % of the channel capacity can be used because of the limiting effects of crosstalk. Most of the research on crosstalk in WDM systems has focused on crosstalk characterization. For example, in [Humblet '90] analysis was used to determine channel spacing required to meet various performance criteria.

A common WDM network topology is the star network. It has the property that all the channels are available to the end user. The crosstalk from undesired channels in this type of network is linear, given reasonable restrictions on signal power and channel spacing (see the section on nonlinear crosstalk in Chapter 4). Grating-based

demultiplexers [Soole '92, Cremer '91] direct the signals to different points along an array of photo-detectors. Other types of demultiplexers share this property [Aisawa '93]. The performance of these demultiplexers is imperfect so the light from a given channel spreads over several adjacent detectors. We propose to simultaneously detect a number of the adjacent detectors, weight them, and combine them with the desired channel. The adaptive algorithm adjusts the weights to cancel the crosstalk in a direct-detection receiver. This technique will reduce the crosstalk in the final filter output. Another network to which this technique will apply is the LARnet recently proposed by Bell Labs [Zirngibl '95].

We note here that some proposed WDM networks will produce crosstalk that post-detection array processing cannot eliminate. These networks have wavelength reuse where a fiber passes through a drop/add node that removes some signals and adds others [Brackett 93]. The drop/add nodes are imperfect, so each time they drop a signal a residue remains. Signal removal may happen several times before the channel terminates in a receiver. The desired signal may have crosstalk from previously removed signals with the same wavelength. Since the residue is at the same wavelength, it can mix with the desired channel to produce nonlinear beating signals in the detector outputs. For this reason it is sometimes referred to as coherent crosstalk. Post-detector array processing cannot remove same-wavelength crosstalk because light from both the desired signal and the undesired crosstalk have the same distribution across the detector array. The distributions are identical even if no nonlinear beating signal arises from the two same-wavelength signals. Other techniques must be employed to eliminate same-wavelength crosstalk. For example, use of a hybrid WDM/CDM network [Foschini '88] to suppress same-wavelength crosstalk by using different codes for any two signals sharing the same wavelength.

Figure 2 shows a block diagram of the proposed array processing receiver. Such a canceller gives several benefits: 1) channels can be packed more closely; 2) less expensive or smaller gratings with larger spot sizes can be used; and 3) looser specifications on laser frequency drift, linewidth, or changes in grating performance because of temperature effects can be tolerated. Adaptive cancellation also mitigates the effects of component degradation because of aging or adverse conditions in the field.

From the above figures and brief descriptions, crosstalk cancellation and channel equalization may seem dissimilar. However, they are essentially the same problem. Consider two idealized matrix equations (noise is neglected). The first is $y=\mathbf{W}^T\mathbf{GS}$. The equation relates the canceller output, y, to the vector of WDM channels, $\mathbf{S}$, through the weight vector, $\mathbf{W}$, and a matrix, $\mathbf{G}$, that represents the nonideal effects of the demultiplexer. The goal is to find $\mathbf{W}$ so that $\mathbf{G}^T\mathbf{W} = \mathbf{I}_d$, a column vector of all zeros except for a one in the d^{th} position, where d is the index of the desired channel. Now consider a matrix version of the equalizer relationship, $y=\mathbf{W}^T\mathbf{Ha}$. This equation relates the equalizer output y to the column vector of data symbols $\mathbf{a}$ through the tap weight vector $\mathbf{W}$ and a matrix $\mathbf{H}$. $\mathbf{H}$ is determined solely by the impulse response of the channel (the channel must be FIR for the matrix to have finite dimensions). The goal is to find $\mathbf{W}$ so that $\mathbf{H}^T\mathbf{W}= \mathbf{I}_n$, a vector of all zeros except for a one in the n^{th} position, where n is an arbitrary delay value. So we see that different delays for the equalizer problem correspond to different channels for the crosstalk canceller problem. The applications are essentially the same problem with the equalizer trying to produce a pure delay and the canceller trying to produce a pure signal. The main difference between the two applications is that the equalizer succeeds if it produces *any* delay while the canceller must produce a *specific* channel.

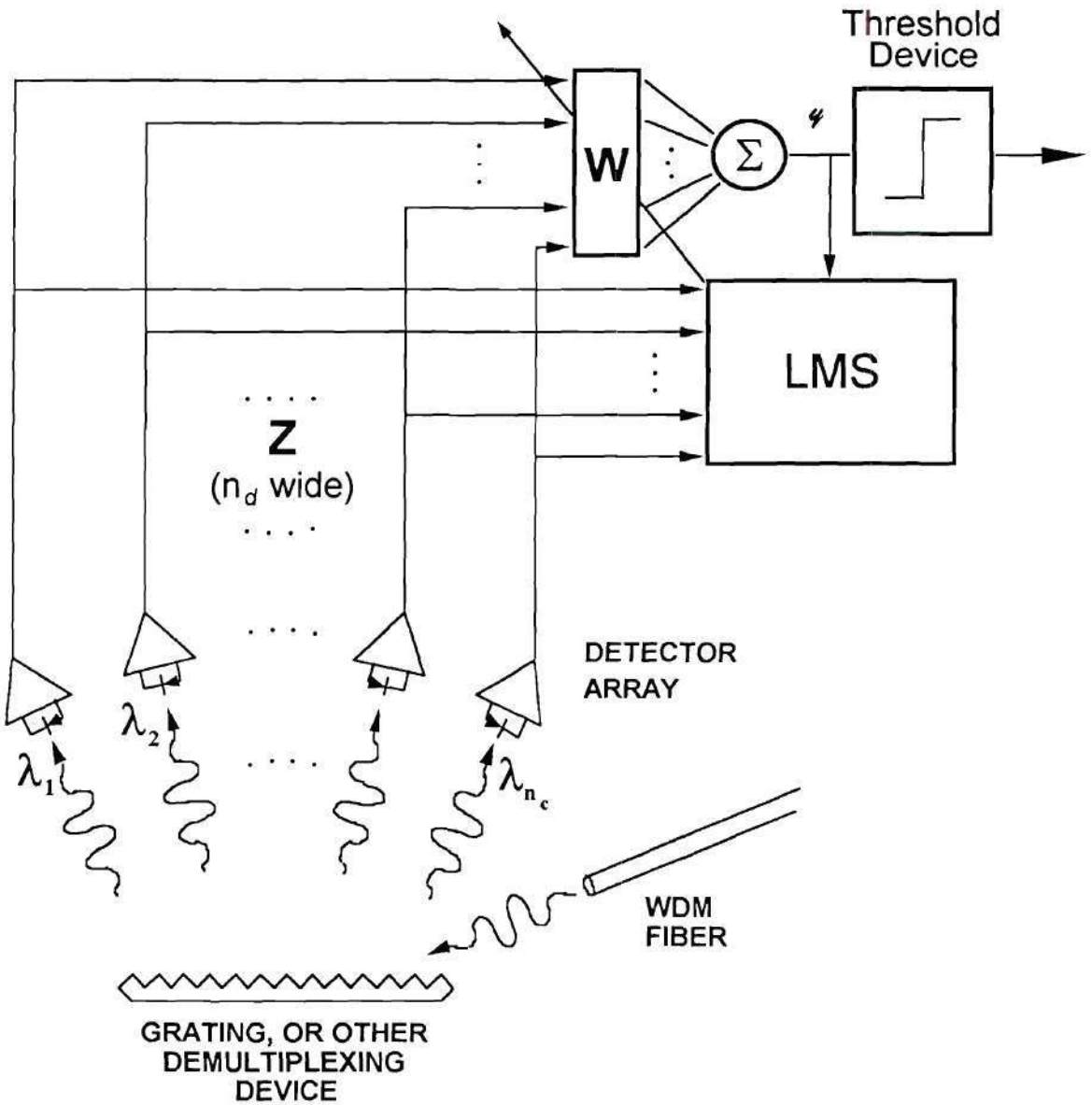


Figure 2: Block diagram of proposed crosstalk cancellation scheme.

We developed four algorithms for crosstalk cancellation. Our simulations show that all four algorithms can more than double the data carrying capacity of a WDM optical link.

We borrowed two algorithms from the field of blind equalization, the Decision Directed Algorithm (DDA) and the Constant Modulus Algorithm (CMA). They require zero mean signals so the receiver must be a.c. coupled to remove the d.c. component of the detector currents. The algorithms are augmented with a weight management technique we call "peak forcing" to ensure that the filter converges to the desired channel.

We also developed two new algorithms. The Constrained-Decision-Directed (CDD) algorithm (previously reported in [Minardi '92,'95]) operates on any digital intensity-modulated signal that includes the absence of light (a "zero") as one of the symbols. It uses a linear constraint and is decision-directed, which means that the flow of the algorithm depends on the symbol decisions of the receiver. The second algorithm, the pilot tone (PT) LMS algorithm [Minardi '95], uses pure sine waves added to the laser drive current, such that each channel has a sine wave (pilot tone) of unique frequency. The receiver generates a sine wave with the proper frequency and phase for use as a desired signal with an unconstrained LMS algorithm. Pilot tones eliminate the need for a linear constraint and decision-directed feedback. The pilot tone LMS algorithm works with analog or digital intensity-modulated data, and the receiver automatically configures itself to account for laser frequency drift. The PT algorithm is not a "blind" technique because it requires cooperation of the transmitter. However, the tasks of data transmission and weight training are multiplexed in frequency so that data transmission is uninterrupted.

This thesis also addresses some of the important non-ideal characteristics of WDM networks that may affect crosstalk cancellation: laser frequency drift and photodetector beating nonlinearities.

The knowledge gained from developing these algorithms led to new insight in the problem of convergence for blind equalizers. Misconvergence has been observed in all equalizers that use a class of techniques known as "Bussgang" algorithms. [Johnson '91] contains a good overview of this problem. We show that Bussgang blind equalizers (for example, the Constant Modulus Algorithm and the Decision Directed Algorithm [Sato '75, Godard '80]) always have stable points with unacceptably high levels of residual ISI. We show that the weights that cause the overall channel-equalizer response to ideally approximate a pure delay (in a mean square error sense) are within the region of convergence of a local minimum of the Bussgang cost function. For any nontrivial channel the approximations for some delay values are poor; usually extremely long and short delays. If a Bussgang algorithm is initialized with these weights it misconverges. Monte-Carlo simulations of two popular algorithms, DDA and CMA, verified this prediction.

This insight led to the development of a new class of blind algorithms. They incorporate linear constraints in weight space and have cost functions based on the variance of the modulus of the equalizer output raised to a power. We performed Monte-Carlo simulations that demonstrate their robust convergence properties.

The thesis is organized as follows. Chapter 2 presents background material and concepts related to the proposed research. Our new work in blind equalization is reported in Chapter 3. Chapter 4 describes our new crosstalk algorithms. Results of WDM crosstalk cancellation simulation are described in Chapter 5. Chapter 6 presents conclusions.

CHAPTER 2.

BACKGROUND AND HISTORICAL PERSPECTIVE

2.1.Other Work in Crosstalk Cancellation

Close channel spacing in dense (WDM) networks causes linear crosstalk in the receivers due to imperfect performance of the demultiplexing elements. In addition to the work already mentioned, other authors have treated digital receivers with multichannel linear crosstalk. [Salz '85] derived an optimal (in the MSE sense) receiver filter for a multichannel digital receiver of BPSK signals. No adaptive algorithms were proposed. Hoenig et al. applied the LMS algorithm to bundles of twisted wire pairs [Hoenig '90]. Aisawa and Hargreaves [Aisawa '93, Hargreaves '93] demonstrated WDM crosstalk cancellers using neural networks that require training sequences. Batra and Barry [Batra '95] have recently demonstrated vector version of the LMS algorithm for multiple-input-multiple-output channels with both ISI and crosstalk. Ho and Kahn [Ho '95] proposed a technique similar to our pilot tone algorithm at OFC '95. To our knowledge, we were the first to propose post detection adaptive crosstalk cancellation. Apart from the above cited work, we have found no other literature on the subject. The general area of adaptive signal processing is, however, large and varied. We will examine two fields that relate to the crosstalk cancellation problem and suggest possible algorithms. The areas are adaptive array processing and blind equalization.

2.2.Problem Description

Figure 2 shows the proposed WDM receiver. Assume that n_c WDM channels are demultiplexed by a grating and then photodetected and amplified. The imperfect

demultiplexer means each detector output is a linear combination of all the channels plus thermal noise. The n_d-length vector $\mathbf{Z}=\mathbf{GS}+\mathbf{n}$ represents the receiver outputs, where $\mathbf{S}$ is an n_c-length vector of optical signal intensities for each channel without crosstalk, $\mathbf{n}$ is an n_d-length vector of receiver thermal noise and $\mathbf{G}$ is a n_c by n_d matrix of crosstalk gains. g_{ij} is the gain of the j^{th} wavelength by the i^{th} direct detection receiver. The i^{th} row of $\mathbf{G}$ describes how each laser couples into detector "i" and the j^{th} column of $\mathbf{G}$ describes how the demultiplexer distributes the light from laser "j" across the detector array. $\mathbf{G}_d$, the d^{th} column of $\mathbf{G}$, is especially important because it represents the light distribution of the desired signal. $\mathbf{G}_c$ is a n_c-1 by n_d matrix consisting of the remaining columns corresponding to the undesired channels. Similarly we split $\mathbf{S}$ into the desired signal s_d and the vector of the undesired signals $\mathbf{S}_c$. Using this notation $\mathbf{Z}=\mathbf{G}_d s_d+\mathbf{G}_c\mathbf{S}_c+\mathbf{n}$, which clearly shows the three components of the detector currents: desired signal, undesired signals and detector noise.

The receiver outputs are weighted and summed to produce the overall system output

$$y = \mathbf{W}^T\mathbf{Z}=\mathbf{W}^T(\mathbf{GS}+\mathbf{n}). \tag{1}$$

The goal is to find a $\mathbf{W}$ that will minimize the bit error rate. To make the problem tractable we use an alternate criterion; maximize the energy ratio of desired-signal-to-crosstalk-and-noise (SCNR). Before proceeding, we define this performance criterion. Given $\mathbf{W}$, we calculate three important quantities. The expected value of the output noise power,

$$P_n = E\{(\mathbf{W}^T\mathbf{n})^2\}= \mathbf{W}^T E\{\mathbf{n}\mathbf{n}^T\}\mathbf{W}= \mathbf{W}^T\sigma^2\mathbf{I}\mathbf{W} = \mathbf{W}^T\mathbf{W}\sigma^2, \tag{2}$$

the expected value of the output crosstalk power,

$$P_C = E\{(\mathbf{W}^T\mathbf{G}_C\mathbf{S}_C)^2\} = \mathbf{W}^T\mathbf{G}_C E\{\mathbf{S}_C\mathbf{S}_C^T\}\mathbf{G}_C^T\mathbf{W} = \mathbf{W}^T\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_C^T\mathbf{W}, \quad (3)$$

and the expected value of the desired channel signal power,

$$P_d = E\{s_d^2\}\mathbf{W}^T\mathbf{G}_d\mathbf{G}_d^T\mathbf{W}, \quad (4)$$

where $E\{\bullet\}$ denotes expectation, $\mathbf{R}_{CC} = E\{\mathbf{S}_C\mathbf{S}_C^T\}$ and σ^2 is the noise power. SCNR is defined by the ratio:

$$SCNR = \frac{P_d}{(P_c + P_n)} = \frac{E\{s_d^2\}\mathbf{W}^T\mathbf{G}_d\mathbf{G}_d^T\mathbf{W}}{\mathbf{W}^T(\mathbf{G}_c\mathbf{R}_{cc}\mathbf{G}_c^T + \sigma^2\mathbf{I})\mathbf{W}}. \quad (5)$$

We desire an optimum weight is to maximize Equation (5). First, without loss of generality, force the numerator of (5) to be a constant. Then, restate the question as: minimize the denominator of (5) subject to a linear constraint $\mathbf{G}_d^T\mathbf{W} = 1$. Geometrically, the constraint means that $\mathbf{W}$ must lie in an n_d-1 dimensional hyperplane (or affine subspace) perpendicular to the constraint vector $\mathbf{G}_d$ and located a distance 1 from the origin.

Using Lagrange multipliers [Frost '72], the solution is

$$\mathbf{W}_{SCNR} = k(\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_C^T + \sigma^2\mathbf{I})^{-1}\mathbf{G}_d, \quad (6)$$

where k is a scalar constant equal to $[\mathbf{G}_d^T(\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_C^T + \sigma^2\mathbf{I})^{-1}\mathbf{G}_d]^{-1}$. If $\mathbf{W}$ varies by a scale factor SCNR is unchanged so the exact value of k is unimportant.

SCNR is difficult to work with because it is a ratio. Mean Square Error (MSE) is commonly used as an alternative to SCNR. MSE is defined as the variance of the difference between the filter output and the desired crosstalk- and noise-free output signal given by:

$$\begin{aligned} J=E\{e^2\}&= E\{(y-s_d)^2\} = E\{(\mathbf{W}^T\mathbf{Z}-s_d)^2\} = E\{(\mathbf{W}^T(\mathbf{GS}+\mathbf{n})-s_d)^2 \\ &= \mathbf{W}^T(\mathbf{G}E\{\mathbf{SS}^T\}\mathbf{G}^T + E\{\mathbf{nn}^T\})\mathbf{W}-2\mathbf{W}^T(\mathbf{G}E\{\mathbf{S}s_d\}+E\{\mathbf{n}s_d\})+E\{s_d^2\} \\ &= \mathbf{W}^T(\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T + \sigma^2\mathbf{I})\mathbf{W} - 2\mathbf{W}^T\mathbf{G}\mathbf{R}_{Sd} + E\{s_d^2\}, \end{aligned} \tag{7}$$

where $\mathbf{R}_{SS}$ is $E\{\mathbf{SS}^T\}$, and $\mathbf{R}_{Sd}$ is $E\{\mathbf{S}s_d\}$, the expected value of the desired signal multiplied with the signals from each channel. It is important to note that $\mathbf{R}_{Sd}$ is equal to a scale multiple of the d^{th} column of $\mathbf{R}_{SS}$. We use the Wiener-Hopf equation to solve for the the weights that minimize J.

$$\mathbf{W}_{MSE} = (\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T+\sigma^2\mathbf{I})^{-1}\mathbf{G}\mathbf{R}_{Sd} = \mathbf{R}_{ZZ}^{-1}\,\mathbf{R}_{Zd}, \tag{8}$$

where $\mathbf{R}_{ZZ} = (\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T+\sigma^2\mathbf{I})$ and $\mathbf{R}_{Zd} = \mathbf{G}\mathbf{R}_{Sd}$. Even though Equations (6) and (8) look very different it has been shown [Monzingo '80] that $\mathbf{W}_{MSE}$ and $\mathbf{W}_{SCNR}$ differ by only a constant if $\mathbf{R}_{SS}$ is diagonal (which is equivalent to saying that the signals are zero mean and uncorrelated from each other). Hence, $\mathbf{W}_{MSE}$ and $\mathbf{W}_{SCNR}$ produce the same SCNR. This condition is true for direct detection optical receivers with a.c. coupling (this would have the effect of turning a sequence of zeroes and ones into a sequence of ± ½'s). We sometimes refer to $\mathbf{W}_{MSE}$ as the MMSE weights.

2.3. Adaptive Filters

Many different algorithms such as least mean squares (LMS) and recursive least squares (RLS) drive adaptive filters. LMS [Widrow '76] is simpler and requires on the order of n_d operations each iteration while RLS requires on the order of n_d^2 operations. The tradeoff is that RLS shows faster convergence than LMS [Treichler '87]. We imagine the canceller to be integrated on the same chip as the detector array and implemented partially with analog electronics [Hirotsu '93]. For this reason we focus on LMS.

2.3.1. Least Mean Squares Algorithm

LMS is based on the gradient search algorithm, also known as the method of steepest descent. The negative gradient of the squared error, $-\nabla_{\mathbf{W}}(E\{e^2\})$, is the direction in $\mathbf{W}$-space that gives the greatest improvement in $E\{e^2\}$. By taking a small step in this direction at each iteration, the weights converge to $\mathbf{W}_{\mathrm{MSE}}$. In general, $\nabla_{\mathbf{W}}(E\{e^2\})$ is not known; however, if $E\{e^2\}$ is replaced by the instantaneous error energy $e^2 = (y - s_d)^2$ then

$$\nabla_{\mathbf{W}}(e^2) = 2e\nabla_{\mathbf{W}}(e) = 2e\nabla_{\mathbf{W}}(\mathbf{W}^T\mathbf{Z}-s_d) = 2e\mathbf{Z}. \qquad (9)$$

LMS uses Equation (9) as a noisy estimate of the gradient. This algorithm is simply:

Form error output: $e = \mathbf{W}_{old}^T\mathbf{Z} - s_d$

Update the weights: $\mathbf{W}_{new} = \mathbf{W}_{old} - \mu e\mathbf{Z}$

Repeat,

where μ is the step size, a small positive real number.

The LMS algorithm uses y and $\mathbf{Z}$ to drive $\mathbf{W}$ towards the optimum value. In many applications s_d is not known, and producing the desired signal is typically the central problem in designing a practical adaptive algorithm. An obvious solution is a training

sequence, where the transmitters periodically send known signals that are prestored in the receiver. The LMS algorithm compares the received training sequence to the stored data and uses the difference to update the weights. This method requires cooperation between the transmitter and receiver and interrupts data transmission. For this reason, we desire algorithms that minimize the cooperation needed for the adaptive processing. In the sequel, we will refer to the standard LMS algorithm as "training-sequence LMS".

2.4. Adaptive Array Processing

Crosstalk cancellation in WDM receivers is an application of array processing. Adaptive RF array processing for radar and communications applications has been an active area of research for many years. This section reviews some relevant results in this area. For further reading we recommend some of the many fine texts in this area, such as: [Monzingo '80, Johnson '93, Haykin '92].

2.4.1. Definitions

Figure 3 shows a narrow band RF uniform linear array (ULA) consisting of m isotropic elements, followed by a weight-and-sum beam former. A set of n narrow band plane wave signals impinge on the array, each from a different angle. The baseband vector of signals $\mathbf{S}(t)$ is complex valued. An m by n matrix $\mathbf{V}$ couples $\mathbf{S}$ into $\mathbf{X}$, the vector of antenna element outputs. The output of the beamformer is therefore

$$s(t) = \mathbf{W}^H\mathbf{X}(t) = \mathbf{W}^H[\mathbf{V}\mathbf{S}(t)+\mathbf{n}(t)]. \tag{10}$$

where $\mathbf{n}(t)$ is a vector of Gaussian receiver noise that is white in both space and time and the superscript "H" denotes the conjugate transpose. Note the similarity to Equation (1), which defines the output of the crosstalk canceller. The assumptions of a uniform linear

array and a plane wave means that the matrix **V** has a particular structure. For a given signal $s_i(t)$ the corresponding column of **V** has the form

$$\mathbf{V}_i = \mathbf{e}(\theta_i) = \left[1,\ e^{j\frac{d}{\lambda}\sin(\theta_i)},\ e^{j2\frac{d}{\lambda}\sin(\theta_i)},\cdots,e^{j(m-1)\frac{d}{\lambda}\sin(\theta_i)}\right]^H. \qquad (11)$$

Note that for a given array geometry, $\mathbf{V}_i$ is a function of a single variable, the angle of arrival θ_i. Similarly a WDM receiver response is a function of a single variable, laser wavelength, albeit in a more complicated form.

2.4.2.Constrained Optimization

In radar applications the desired signal is competing with unintentional interference from other sources, hostile jamming signals, and returns from large scatterers in the radar sidelobes. The desired signal, the radar return, is not available as an input to an adaptive algorithm. The radar community handles this problem by the use of a constraint. They suppose that the desired signal arrives from a certain direction and use a linear constraint to force the array pattern to have a high gain in that direction. Next, minimize the energy by adjusting weights that are restricted to the space orthogonal to the constraint vector. If a signal arrives from the test direction it will pass through to the array output while all other signals are suppressed. Formally stated, the problem is to minimize the output power of the adaptive beamformer $E\{|s(t)|^2\}$ subject to the constraint $\mathbf{C}^H\mathbf{W} = c$: $\min(\mathbf{W})E\{|s(t)|^2\}$ $= \mathbf{W}^H(\mathbf{V}\mathbf{R}_{SS}\mathbf{V}^H + \sigma_n{}^2\mathbf{I})\mathbf{W} = \mathbf{W}^H\mathbf{R}_{XX}\mathbf{W}$, subject to $\mathbf{C}^H\mathbf{W} = c$.

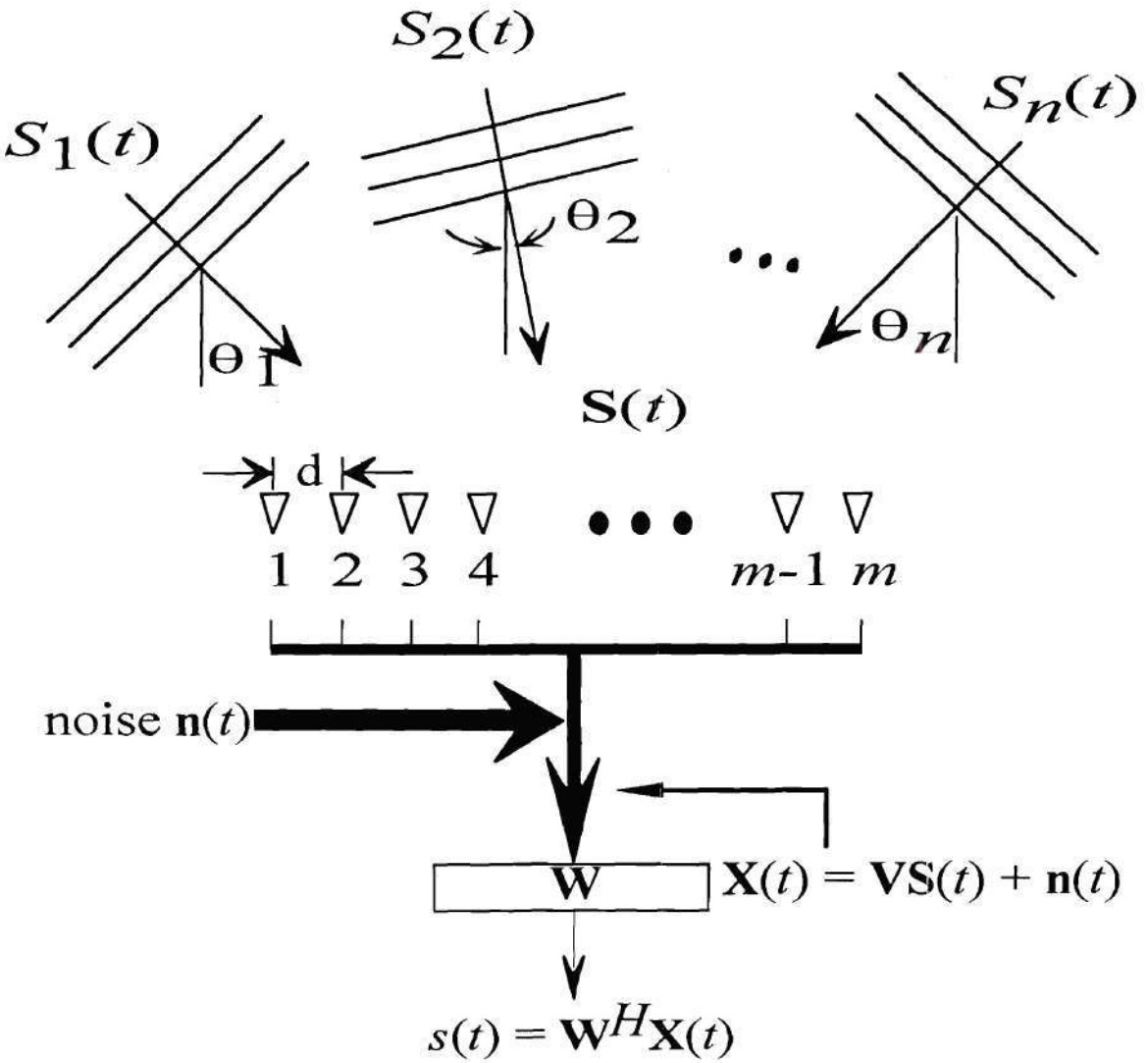


Figure 3: Uniform linear array.

The solution is [Monzingo '80]

$$\mathbf{W}_C=\mathbf{R}_{XX}^{-1}\mathbf{C}(\mathbf{C}^H\mathbf{R}_{XX}^{-1}\mathbf{C})^{-1}c. \tag{12}$$

If $\mathbf{C} = \mathbf{V}_i$ (i.e., it matches a column of $\mathbf{V}$) and if $\mathbf{R}_{SS}$ is diagonal, then the weights are optimum for selecting out $s_i(t)$ from the competing signals in both an MSE and SCNR sense (see Equations (8) and (6)). The fact that $\mathbf{V}_i$ is indexed by a single variable, θ_i, makes searching all possible $\mathbf{V}_i$ practical. These ideas are central to the Minimum Variance and eigenanalysis techniques described below.

2.4.2.1.Minimum Variance Beamformer

This technique is a straightforward application of the above ideas. The constrained array output power is $\mathbf{W}_C{}^H\mathbf{R}_{XX}\mathbf{W}_C = (\mathbf{C}^H\mathbf{R}_{XX}{}^{-1}\mathbf{C})^{-1}c2$ [Monzingo '80]. The array power output is calculated for each look direction, $\mathbf{e}(\theta)$,

$$P_{MV}(\theta) = (\mathbf{e}(\theta)^H\mathbf{R}_{XX}{}^{-1}\mathbf{e}(\theta))^{-1}. \tag{13}$$

The quantity $P_{MV}(\theta)$ forms sharp peaks whenever θ matches one of the signal directions (as long as the signals are not too close together). The peak locations identify the signal directions then the weights are determined by Equation (12) [Johnson '93]. The term "minimum variance" refers to the minimization of the output power, which in the case of random signal is the variance.

2.4.2.2.Eigenanalysis Techniques

These techniques require the determination of the eigenvectors and eigenvalues of $\mathbf{R}_{XX}$. If the number of signals, n, is less than the number of sensors in the array, m, then

the space spanned by the eigenvectors can be broken up into two orthogonal subspaces, the signal subspace with dimension n and the noise subspace with dimension m-n. The eigenvectors associated with the m-n smallest eigenvalues form a basis for the noise subspace. Several algorithms identify the noise subspace and use it to determine signal locations. For example, the MUltiple-SIgnal-Classification algorithm (MUSIC) [Schmidt '86] modifies the minimum variance power expression (Equation (13)) by replacing $\mathbf{R}_{XX}$ with $\mathbf{R}_{MUSIC}$ where

$$\mathbf{R}_{MUSIC} = \mathbf{V}_n \mathbf{V}_n^H \tag{14}$$

and $\mathbf{V}_n$ is an m by (m-n) matrix where each column is an eigenvector associated with the noise subspace. A power-like quantity is then calculated (again, note the similarity to Equation (13)),

$$P_{MUSIC}(\theta) = (\mathbf{e}(\theta)^H \mathbf{R}_{MUSIC}^{-1} \mathbf{e}(\theta))^{-1}. \tag{15}$$

When θ matches a signal direction then $\mathbf{e}(\theta)$ is orthogonal to all the columns of $\mathbf{V}_n$ and P_{MUSIC} forms a peak. MUSIC does a better job of separating closely spaced signals than the minimum variance technique at the expense of more calculations. It also requires that the number of signals be strictly less than the number of array elements.

2.4.2.3.Generalized Sidelobe Canceller

MV and MUSIC are so-called block adaptive techniques. The correlation matrix $\mathbf{R}_{XX}$ is estimated from a block of received data and used to identify the signal directions and form the weights. The generalized sidelobe canceller (GSC) is a receiver architecture that allows continuous adaptation to the constrained optimization problem [Griffiths '82].

Continuous adaptation lets the array adapt to a nonstationary environment. Figure 4 shows a block diagram of the GSC. The upper branch forces the weights to satisfy the constraint **C**. It is fixed and not adaptive. The lower branch is the adaptive portion, the signal vector passes through the "blocking matrix" **B**. The blocking matrix reduces the dimension of the signal by one and must have the constraint vector **C** span its null space (i.e., $\mathbf{BC} = \mathbf{0}$). A suitable adaptive algorithm adjusts the adaptive weights $\mathbf{W_a}$ to minimize the total array output power $E\left\{\left|\left(\mathbf{e}(\theta)^H - \mathbf{W_a B}\right)\mathbf{X}\right|^2\right\}$. If θ matches a signal then that signal cannot appear in the bottom branch and the weights cannot act on it. Instead, the weights configure themselves to annihilate the remaining signals. Our crosstalk cancellation and blind equalization algorithm simulations that employ linear constraints sometimes use the GSC architecture [Minardi '93].

2.5.Blind Equalization

Equalization is the technique of filtering the received signal to reduce inter-symbol interference (ISI). Equalization is required to achieve reliable performance in many high speed digital communications applications. Adaptive equalization uses a digital filter with adjustable weights to accommodate unknown channels or slowly time varying channels. As we showed in Chapter 1, equalization in digital receivers is similar to the problem of crosstalk cancellation, with ISI analogous crosstalk. Early adaptive schemes required the use of a training sequence that is prestored in the receiver.

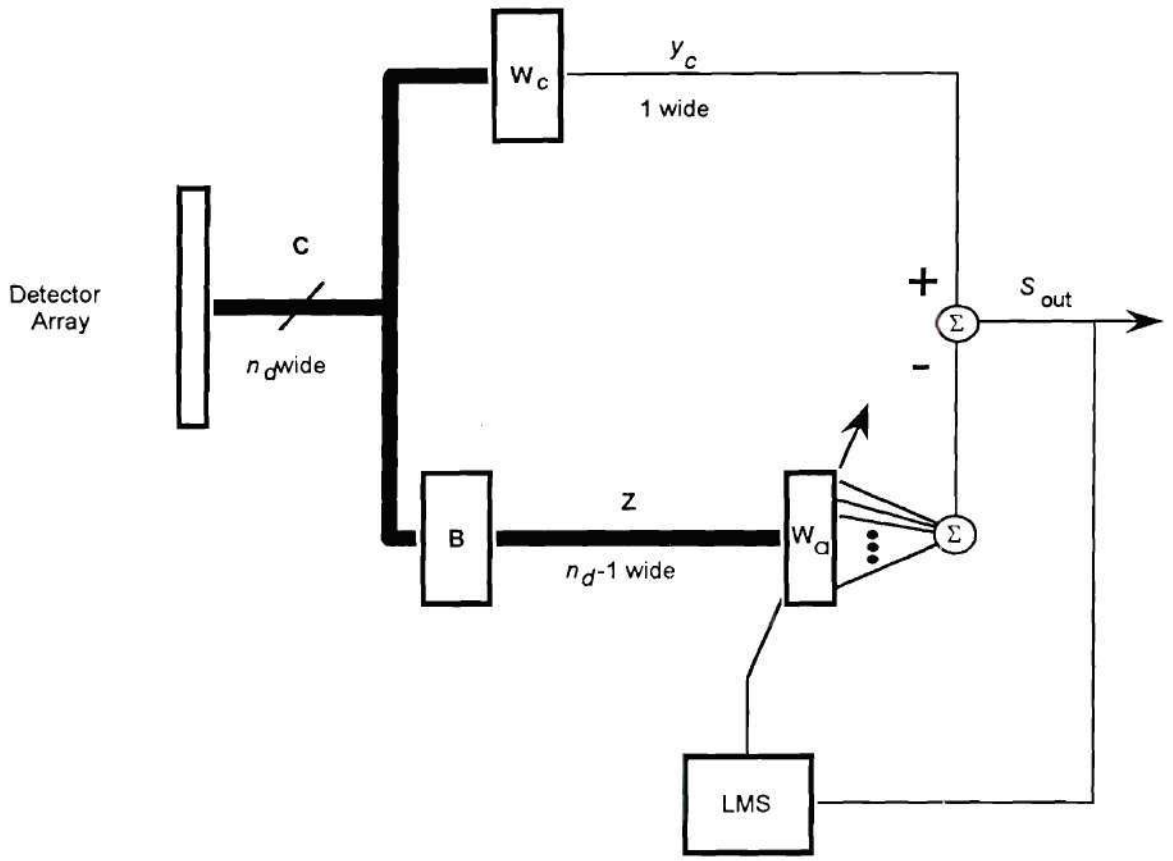


Figure 4: Block diagram of the Generalized Sidelobe Canceller (GSC).

Training interrupts data transmission so in many cases the transmission of a training sequence is impractical. For example, in a multipoint network, the transmitter needs to interrupt data transmission to train a new receiver when it is activated. In multipath fading channels the receiver may require frequent retraining to adapt to a time varying channel. This problem has led to the search for blind adaptive algorithms that determine the proper weights solely by examining the channel output, relieving the network of the overhead due to training sequences.

The following Theorem from [Benveniste '80] indicates how it is possible to identify a channel solely from the channel output and knowledge of the statistics of the channel input:

> **Theorem:** Let a_m pass through a linear time invariant channel. The symbols of the sequence a_m is independent and identically distributed (i.i.d.), with a non-Gaussian, zero mean and symmetric distribution. If the output of a linear equalizer y_m has the same distribution as the channel input a_m then, $y_m = \pm a_{m-k}$ and the sequence has been recovered except for an arbitrary fixed delay and arithmetic sign.

In other words, matching the statistics of the equalizer output to the known distribution of the input recovers the input. Shalvi and Weinstein [Shalvi '90] showed that only the second and fourth moments of the output need be made equal to the second and fourth order moments of the input symbols to recover the input sequence.

2.5.1.Definitions

Figure 1 shows a block diagram of a channel and adaptive equalizer. A transmitter generates a sequence of symbols, a_m, from a discrete signal constellation (usually pulse amplitude modulation (PAM) or quadrature amplitude modulation (QAM)) and transmits it through a linear time-invariant channel. The a_m are i.i.d. random variables with zero

mean, variance one. The probability density function $f_a(x)$, is symmetric and assumed to be known to the receiver. The channel model includes transmitter and receiver pulse shaping filters and has impulse response h_m and transfer function $H(\omega)$. We assume that the receiver is synchronized with the transmitter and the channel output is sampled once per symbol period. We call the channel output sequence x_m,

$$x_m = a_m * h_m = \sum_{k=-\infty}^{\infty} a_{m-k} h_k \,. \tag{16}$$

The channel output is combined with a zero mean additive white Gaussian noise (AWGN) sequence n_m (with variance σ_n^2) to create a received sequence that is passed through an adaptive equalizer with impulse response w_m to give the equalized output y_m,

$$y_m = (a_m * h_m + n_m) * w_m = (x_m + n_m) * w_m = \sum_{k=0}^{N-1} (x_{m-k} + n_{m-k}) w_k \,, \tag{17}$$

where we have assumed that the equalizer is a transversal finite impulse response (FIR) filter with N taps. A third sequence often used for analysis purposes is the combined channel-equalizer impulse response $t_m = w_m * h_m$. Let h_m be FIR with length M, so that an alternative matrix expression relating a_m and y_m is

$$y_m = \mathbf{W}^T(\mathbf{H}\mathbf{a}_m + \mathbf{n}_m) = \mathbf{T}^T\mathbf{a}_m + \mathbf{W}^T\mathbf{n}_m, \tag{18}$$

where $\mathbf{T} = [t_0, \cdots, t_{N+M-2}]^T$, $\mathbf{W} = [w_0, \cdots, w_{N-1}]^T$, $\mathbf{a}_m = [a_m, \cdots, a_{m-N-M+2}]^T$ and $\mathbf{H}$ has N rows and $N+M-1$ columns.For example, for $M=4$ and $N=4$, $\mathbf{H}$ is:

$$\mathbf{H} = \begin{pmatrix} h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 \\ 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 \\ 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 \\ 0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 \end{pmatrix}. \tag{19}$$

H is a matrix of coupling factors where h_{ij} is the gain of the j^{th} element of symbol vector $\mathbf{a}_m$ into the i^{th} element of $\mathbf{x}_m$. The i^{th} row of **H** describes how each symbol couples into element "i" of $\mathbf{x}_m$ and the j^{th} column of **H** describes how symbol "j" is distributed throughout $\mathbf{x}_m$. **H** always has more columns than rows so the system is, by nature, undetermined, which means that some residual ISI will always remain. The optimum weights that minimize the MSE for various delays are the columns of $(\mathbf{H}\mathbf{H}^T+\sigma_n^2)^{-1}\mathbf{H}$. Note that the weights are independent of the signal constellation. If noise is neglected this becomes $(\mathbf{H}^T)^{\#}$, where "#" denotes the Penrose pseudo-inverse. The m^{th} column of $(\mathbf{H}^T)^{\#}$ are the weights that produce an overall system response, t_m, that is the best (in an MSE sense) approximation to a pure delay of m samples. Column 0 gives a delay of 0, column 1 a delay of 1 and so on (note that we label the first column 0 so the column and delay indices match).

2.5.2.Methods Using Higher Order Statistics (HOS)

Classical methods that minimize the MSE, such as LMS, use only second order statistics of the channel and equalizer output, for example, the autocorrelation or power spectral density. Second order statistics are phase blind and only contain information about $|H(\omega)|$ [Ding '94]. So standard least squares type methods cannot identify channels with unknown phase. Higher order statistics (HOS) of the channel output contain information about both the magnitude and phase of the channel and can be used for blind equalization. Two classes of HOS algorithms have emerged. The Bussgang algorithms implicitly use

HOS by using a memoryless nonlinear function of the channel output to drive an LMS-type stochastic gradient algorithm [Sato '75, Bellini '86, Godard '80]. Polyspectral algorithms explicitly use HOS. Estimates of (usually) second and fourth order statistics are generated and used to identify h_m [Tong '94, Hatzinakos '91, Pan '88].

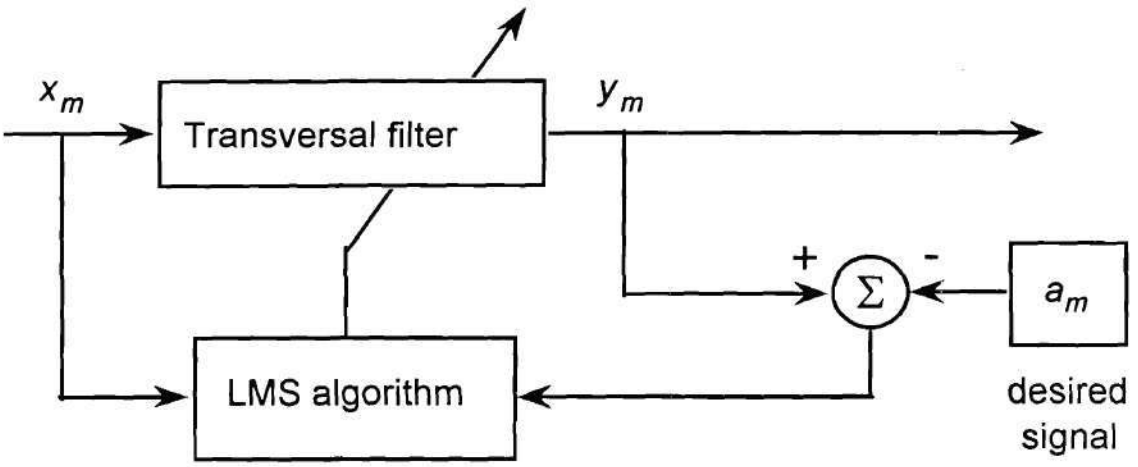
2.5.2.1.Implicit use of HOS, Bussgang Algorithms

These algorithms are extensions of the training-sequence LMS algorithm. In all cases the training sequence is replaced by the output of a memoryless nonlinear function of y_m, $g(y_m)$. Figure 5 gives a simple block diagram of the standard LMS receiver and a Bussgang type of receiver. The input sequence is assumed to be an i.i.d. sequence of symbols from some known non-Gaussian distribution. We will also assume that a_m is real for simplicity. All the algorithms discussed vary only in the form of $g(\bullet)$. The output of the nonlinear element is fed back to update the tap weights using an LMS type algorithm:

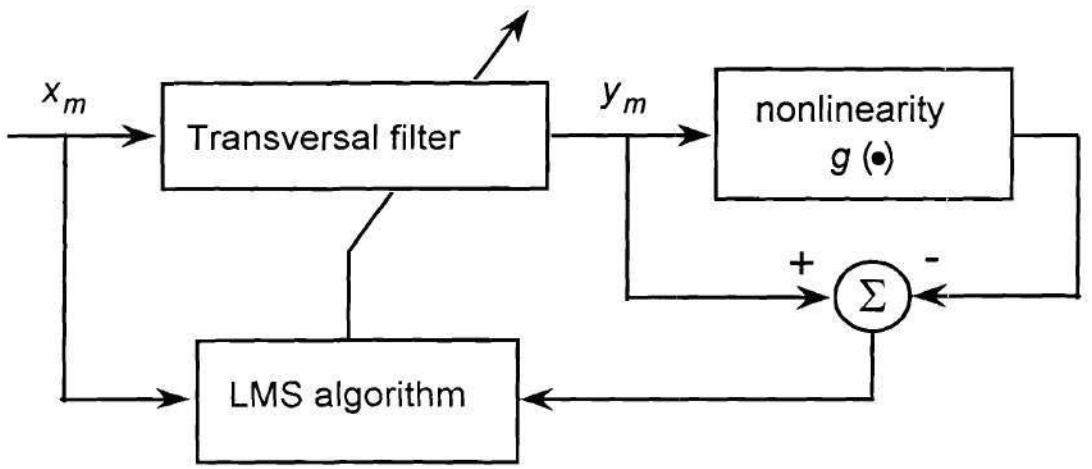
$$\mathbf{W}_{m+1} = \mathbf{W}_m - \mu \mathbf{x}_m \left(y_m - g(y_m) \right). \tag{20}$$

Continuing our comparison to LMS we see that $g(y_m)$ replaces the desired signal and $e_m = y_m - g(y_m)$ fulfills the role of the error. Finally, we define an underlining cost function, $J(\mathbf{W})$, that is to be minimized. It is determined by integrating $E\{\mathbf{x}_m(y_m - g(y_m))\}$ with respect to $\mathbf{W}$. Unlike the training-sequence LMS case, the cost function J has more than one local minimum.

The algorithm converges when $E\{\mathbf{W}_{m+1}\} = E\{\mathbf{W}_m\}$ for all m greater than some integer k and $\mathbf{W}_o = E\{\mathbf{W}_m\}$ is a stable point of the algorithm. A necessary condition for stability is that the expected value of the update term in Equation **(20)** must equal **0**:



a) Training sequence LMS equalization



b) Bussgang blind equalization

Figure 5: Comparison of training sequence and bussgang type equalizers.

$$E\{\mathbf{X}_m(y_m - g(y_m))\} = \mathbf{0}. \tag{21}$$

$$E\{x_{m-k}g(y_m)\} = E\{x_{m-k}y_m\},\ k = 0,1,\ldots,N. \tag{22}$$

If the weights have converged (m is large) and if we assume that the number of taps, N, is large then the N equations defined above can be shown [Haykin '91] to satisfy approximately:

$$E\{y_m y_{m-k}\} = E\{g(y_m)y_{m-k}\}. \tag{23}$$

A stochastic process satisfying Equation (23) is called a Bussgang process [Bussgang '52], Successful convergence makes y_m a Bussgang process hence the name Bussgang algorithm.

Finally, we note that by replacing y_m with $\mathbf{W}^T\mathbf{x}_m$ and expressing $g(\bullet)$ as a power series we see that algorithm convergence imposes relationships on the HOS of x_m:

$$\begin{aligned} E\{\mathbf{x}_m(y_m - g(y_m))\} &= E\left\{\mathbf{x}_m\left(-g(0) + y_m - g'(0)y_m - \tfrac{1}{2}g''(0)y_m^2 - \tfrac{1}{6}g'''(0)y_m^3 + \ldots\right)\right\} \\ &= \left(1 - g'(0)\right)\mathbf{R}_{\mathbf{xx}}\mathbf{W} - \tfrac{1}{6}g'''(0)E\left\{\mathbf{x}\mathbf{x}^T\mathbf{W}\mathbf{W}^T\mathbf{x}\mathbf{x}^T\right\}\mathbf{W} + \ldots = \mathbf{0} \end{aligned} \tag{24}$$

2.5.2.1.1.Decision Directed/Minimum Mean Square Error (MMSE)

If a training signal is not available we can estimate the symbol a_m from the output and use it as an ersatz training sequence. Sometimes the literature does not consider this a Bussgang technique, or even a blind technique. However, [Haykin '91] showed that it satisfied the conditions required for the former and it is obviously the latter. The

memoryless nonlinearity $g(\bullet)$ is the output of the decision circuitry, for example if $a_m = \pm 1$, $g(y_m) = \text{sgn}(y_m)$ (sgn(•) is the sign function). Haykin reported as a rule of thumb that the decision directed technique will not converge if the error probability exceeds 10^{-2}. For this reason it cannot be used to initialize a filter in the face of severe ISI but works well for an equalizer that has already converged adapting to slow changes in the channel. In fact, the goal of most blind equalizers is to equalize the channel to a point where the eye is partially opened and the algorithm can convert to a decision directed algorithm [Haykin '91].

2.5.2.1.2.Sato algorithm

Sato was the first to propose blind equalization for m-ary PAM systems [Sato '75]. The nonlinear function is

$$g(y) = \gamma \text{sgn}(y), \text{ where } \gamma = \frac{E\{x_m^2\}}{E\{|x_m|\}}. \tag{25}$$

The associated cost function is

$$J=E\{(|y|-\gamma)^2\}. \tag{26}$$

The Sato algorithm considers a multilevel PAM signal to be a binary signal and treats all the additional bits of information as additive noise. Note that if the signal is binary, the Sato and decision directed algorithms are equivalent. This algorithm has better convergence properties than the decision directed algorithm although it converges more slowly and has a large excess MSE in the steady state due to the inaccuracy of the error signal [Shynk '91].

2.5.2.1.3.Combined Sato-Decision Directed

Several algorithms combine Sato and decision directed algorithms in order get the robust convergence property of Sato and the fast convergence and good steady state features of DDA [Weerackody '90, Benveniste '84].

An example is the stop-and-go algorithm, proposed by Picchi and Prati [Picchi '87]. The idea is simple; at each iteration compute both DDA and Sato errors and then compare their sign. If they have a different sign do not update the weights, i.e., "coast" during that iteration. If the signs agree then update the tap weights using the DDA error. The motivation is that if the signs of the error disagree it is likely that the estimate made by the decision circuitry is unreliable and the decision should not be fed back to the LMS algorithm.

As the algorithm converges and ISI is reduced the flag will tend to stay "on" all the time so we see that the stop-and-go will smoothly transition to a full decision directed mode automatically. [Hatzinakos '91] also presented several modifications of the stop-and-go algorithm..

2.5.2.1.4.Godard/Constant Modulus

[Godard '80] proposed a class of cost functions specifically designed for use with two dimensional signal constellations like QAM and M-ary phase shift keying. He called his cost functions "dispersion" functions and defined them as

$$J_p = E\{(|y_m|^p - R_p)^2\},\ \ R_p = \frac{E\{|a_m|^{2p}\}}{E\{|a_m|^p\}}. \tag{27}$$

The constant R_p scales the weight vector **W** and only affects the overall gain of the filter. If you replace R_p by cR_p (c>0) then a local minimum of the cost function J, located at

$\mathbf{W}_o$, moves to $c^{1/p}\mathbf{W}_o$. Thus, the output SCNR is unaffected. Godard suggested p = 1 or 2 for practical applications, due to the increased dynamic range and sensitivity to the stepsize μ with increasing p. Triechler and Agee [Treichler '85] independently derived the algorithm for the p = 2 case for the equalization of wireless signals in the presence of multipath and interfering signals. Their algorithm is called the constant modulus algorithm (CMA). CMA is the most widely investigated and implemented blind adaptive algorithm [Haykin '91]. For a real signal, the p=1 case is equivalent to the Sato algorithm.

We derive the update equation by taking the gradient of the cost function with respect to $\mathbf{W}$ and substituting the instantaneous gradient estimate. For CMA (p=2) and R_2=1 the update equation becomes

$$\mathbf{W}_{m+1} = \mathbf{W}_m - \mu \mathbf{x}_m (y_m^3 - y_m). \tag{28}$$

Using the definition of e_m let us determine $g(\bullet)$ for CMA

$$g(y_m) = y_m + e_m = 2y_m - y_m^3. \tag{29}$$

Perhaps surprisingly, CMA works well with signal constellations that do not have a constant modulus, such as multilevel PAM and QAM. Godard showed that the cost function has an absolute minimum resulting in zero ISI when the normalized kurtosis of the symbol constellation is less that two, Normalized kurtosis is defined as:

$$\frac{E\{|a_m|^4\}}{E\{|a_m|^2\}^2}. \tag{30}$$

We calculated the normalized kurtosis for all QAM constellations of up to 2500 elements (50 by 50) and PAM constellations of up to 50 elements. The normalized kurtosis is less that 1.4 and 1.8 for QAM and PAM, respectively.

2.5.2.1.5.Conditional Mean

[Bellini '86, '88] approached the equalization problem by attempting to find an "optimal" $g(\bullet)$ for multilevel QAM and PAM signals. He started by expressing the equalizer output as $y_m=c_O a_m+z_m$, where a_m is the true symbol, z_m is the "convolutional noise" due to the residual ISI not removed by the equalizer and $c_O = \sqrt{1-\sigma_z^2}$. Bellini required c_O because he assumed that an automatic gain control maintains the variance of y_m equal to one). The convolutional noise is assumed to be independent of a_m and by the central limit theorem to have a Gaussian distribution (this is a bad assumption if the ISI is dominated by a few symbols and/or the distribution of a_m is very non-Gaussian, like BPSK. However as the filter converges the dominant terms are the first to be suppressed, which helps the Gaussian assumption). Bellini proposes a conditional mean estimator $E\{a_m|y_m\}$ as a suitable $g(\bullet)$. Using the above assumptions the estimator has the form:

$$g(y) = E\{a|y\} = \frac{\sigma_n}{c_0}\frac{Z(y+c_0\sqrt{3}) - Z(y-c_0\sqrt{3})}{Q(y-c_0\sqrt{3}) - Q(y+c_0\sqrt{3})} + \frac{y}{c_0}, \qquad (31)$$

where $Z(y)$ is the normalized Gaussian probability density and $Q(y) = \int_y^\infty Z(w)dw$.

Bellini evaluated and plotted Equation (31) for different PAM constellations and convolutional noise levels. He noticed that plots of Equation (31) do not much vary for the number of PAM levels and for low SCNR levels (less than 10 dB). For high SCNR levels the curve approaches that of a m-ary decision function. Therefore, he reasoned that

a single curve should suffice for initial convergence and opening the eye. The function can be stored in a look-up table. A sigmoid nonlinearity has been suggested as a good fit to Equation (31) [Haykin '91];

$$g(y) = 1.945 \tanh\left(\frac{1.25y}{2}\right). \tag{32}$$

2.5.2.1.6.Convergence Time Comparisons

Shynk et al. [Shynk '91] simulated seven Bussgang techniques and compared their convergence properties. All Bussgang algorithms exhibit a trade-off between convergence speed and steady state MSE. Convergence time decreases with increasing stepsize while steady state MSE decreases. He ran simulations for each technique with several different step sizes, and the convergence time was plotted against the steady state MSE. The plots form a figure of merit that combines the two essential characteristics of adaptive algorithms. In addition, two cases were considered, one with no carrier frequency error and one with a one-cycle-per-200-sample error. The channel was obtained from measurements of a 22.5 Mbaud digital microwave channel in an urban setting and the equalizer had 33 $T/2$ spaced taps. The input sequence was 64-QAM.

When no carrier frequency error is present the dual-mode Sato algorithms perform best. Once the eye is opened they transition to decision directed mode and rapidly converge to a low error. The CMA type algorithms had poorer performance, however, when a frequency error is included CMA algorithms perform best. Because they are completely independent of phase their performance does not degrade. The Sato based algorithms are very degraded and are never able to transition to decision directed mode. This study indicates that the choice of algorithm should be influenced by whether carrier

synchronization occurs before or after the equalizer. Note that the convergence times are always several thousand symbols long, much longer than when adapting with training sequences.

2.5.2.2.Explicit Use of HOS, Polyspectral Algorithms

These techniques estimate the higher order statistics of the channel output and use the estimates plus the known HOS of the input sequence. The statistics are used to generate a set of equations that has a solution that identifies the channel. The HOS are known as "cumulants" and their Fourier transforms are known as polyspectra. Large numbers of calculations and data are required per iteration to make a reasonable estimate of the cumulants. The first explicit HOS method based on polyspecstra is the Tricepstrum Equalization Algorithm (TEA) proposed by Hatzinakos and Nikias [Hatzinakos '91].

2.5.3.Fractional Spacing/ Cyclostationary algorithms

Fractional spacing algorithms were first proposed by Tong, Xu and Kailath [Tong '91] and Gardner [Gardner '91, '91A]. The idea represents a different approach that does not require the use of higher order statistics. Tong et al. noted that previous treatments of the blind equalization problem always assume that the channel input and output were stationary processes. They are stationary under the usual assumptions, one of which is that the received signal is sampled at the symbol rate T. However the underlying continuous time process is not stationary, it is *cyclostationary* at the symbol period T. This means that all of the statistics at time t_0 are repeated at times t_0+nT (n is an integer).

If the equalizer samples at a shorter period $\Delta=T/m$ then a cyclostationary discrete time sequence is created. Using such a sequence [Tong '91] showed that a non-minimum phase channel can be identified using only second order statistics of the cyclostationary channel output. Because it takes fewer samples to estimate second order statistics

fractional spaced techniques converge much faster (by three orders of magnitude! [Tong '91]) than HOS based techniques. Also the probability distribution of the source sequence is unrestricted and can even be Gaussian. Of course the cost is that the sampling hardware must operate m times faster, which may be a problem at very high baud rates.

2.5.4.Convergence and Misconvergence

The following theorem from [Benveniste '80] gives conditions under which an equalizer with an infinite number of tap weights is guaranteed to converge to a pure delay with all ISI eliminated.

> **Benveniste-Goursat-Ruget (BGR) Theorem**: If the density function of a_n, $f_a(x)$, is *sub-Gaussian*, and if $e(y)$ is odd, twice differentiable (except at the origin), $e(0+)\leq 0$ and $e''(y) \geq 0$ for $y \in (0,\infty)$ (at least one of the inequalities must be strict), then stable minima of the cost function exist only when the channel is perfectly inverted by the equalizer (with an arbitrary delay and sign).
>
> *Definition*: sub-Gaussian means that $f_a(x)=Ke^{-f(x)}$ such that $f(x)$ is an even function, $f(x)$ is strictly increasing and $f'(x)/x$ is strictly decreasing for $x \in (0,\infty)$. For example, $Ke^{-|x|^\gamma}$ is sub-Gaussian for $\gamma > 2$. The uniform distribution is sub-Gaussian. (Multilevel PAM is not sub-Gaussian but if it has many levels it is approximated as uniform in the literature [Bellini '88]).

The theorem does not guarantee this property for a realizable equalizer with a finite number of tap weights. [Ding '91] claimed that finite tap weights add many new minima to the cost function J, because the channel convolution matrix $\mathbf{H}$, defined in Equation (18), is underdetermined. Ding noted that a necessary condition for stability is that the expected value of the update term of the Bussgang algorithm be equal to zero,

$$E\{\mathbf{x}_m(y_m - g(y_m))\} = \mathbf{H}E\{\mathbf{a}_m(y_m - g(y_m))\} = \mathbf{0}. \quad (33)$$

Since **H** is underdetermined this can happen in two ways: first, if the expected value portion is **0**, and second, if the expected value is non-zero but still in the null space of **H**. Furthermore, and this point is missed in [Ding '91], the first case can only happen with a finite weight equalizer if the channel is auto-regressive (contains only poles). Proof: The BGR theorem states that the expected value term in Equation (33) can only equal **0** if ISI is completely eliminated. However, it is impossible to completely eliminate ISI with an FIR filter if the channel frequency response contains any zeros.

The guarantee of global convergence for finite length blind Bussgang type equalizers remains an open problem in the literature [Haykin '91, Johnson '91, Verdu '84]. Chapter 3 will have more discussion on this subject.

CHAPTER 3.

ENHANCING CONVERGENCE IN BLIND EQUALIZERS

3.1.Introduction

Bussgang-type blind equalizers have well publicized problems with misconvergence. Misconvergence has been observed in all the popular algorithms. It occurs when the adaptive weights converge to a local minimum of the cost function and leave excessive ISI. For example, [Ding '91] demonstrated misconvergence in CMA using a simulation of a 2-tap FIR equalizer and a one pole AR channel. [Johnson '91] demonstrated misconvergence for five different Bussgang algorithms using a similar channel/equalizer example.

Most theoretical work on the convergence properties of Bussgang equalizers assume an infinite number of tap weights for tractability [Benveniste '80]. For example, [Foschini '85] claims global convergence of the CMA algorithm using arguments based on an infinite tap equalizer. Conclusions made for infinite tap equalizers cannot be extended to realizable equalizers with finite taps. [Ding '91] suggests that misconvergence is possible because **H** (from Equation (18)) is always underdetermined and has a nontrivial null space; however, he gave no method for finding the bad points for realistic equalizers with many taps. He demonstrated misconvergence for a 2-tap equalizer and a single pole channel. The bad point seemed to correspond to a delay of 2 samples instead of one. [Johnson '91] also noted this misconvergence phenomenon. Our simulations have shown the same types of misconvergence for more realistic examples with equalizers with many taps and channels with impulse responses of greater length.

Researchers still look for a nonlinear function, $g(\bullet)$, which will allow convergence to the channel inverse for any channel and initialization. [Verdu '84] states: "To date no such function is known to result in global convergence to the inverse of the channel when the input consists of binary data." [Johnson '91] reiterates this statement.

Our research shows that no memoryless $g(\bullet)$ can be found such that the corresponding Bussgang algorithm will exhibit global convergence. Furthermore, we have a simple method that locates the bad stable points: Initialize the algorithm with columns of the pseudo-inverse $(\mathbf{H}^T)^{\#}$ and let it stabilize. Some of the points will exhibit the poor performance observed in practice. This knowledge, plus insights gained from our algorithms developed for crosstalk cancellation prompted, the development of a new class of blind adaptive algorithms.

The "Variance-Constraint" (VC) algorithms have a channel gain estimate embedded in their cost functions and utilize linear constraints in the weight space. Although the cost functions still have undesirable local minima, the effects of the constraint pushes the minima farther away in weight. Thus, the VC algorithms have robust convergence properties.

3.1.1. The relationship between MMSE weights for pure delay and stable points

As noted in Chapter 2, if we neglect noise, the MMSE weights for a delay of d samples are the d^{th} column of $(\mathbf{H}^T)^{\#}$, $\mathbf{W}_{\text{MSE}}^{(d)} = (\mathbf{H}^T)^{\#}\mathbf{I}_d$ (as defined in Chapter 1, $\mathbf{I}_d$ is the d^{th} column of $\mathbf{I}$). Some of the approximations (usually the first or last few columns) are poor and leave a large amount of residual ISI. These points are within the region of convergence of stable points of the Bussgang equalizer with poor performance. The columns of $(\mathbf{H}^T)^{\#}$ that produce good approximations to a pure delay with little ISI are nearly coincident with stable points with good performance.

We will show that if the signal-to-interference ratio (SIR) of an MMSE weight is high, $\mathbf{W}_{MSE}$ will be a stable point of the algorithm. A low SIR MMSE weight will cause a bias term to arise in the weight update equation that drives the weights away from the MMSE weights. If the bias term is not too great, the algorithm settles to a point that still can be considered an approximation to the pure delay; i.e., the MMSE point is within the region of convergence of a bad point. These bad points are the location of at least some of the types of misconvergences observed in the literature [Johnson '91, Ding '91].

Figure 6 and Figure 7 show contour plots of the cost functions for the two most popular blind algorithms, DDA and CMA. Equation (27) defines the cost functions, with R_p=1 and p=1 for DDA and 2 for CMA. The examples are for a simple 2-element FIR channel (h_m={1.0, 0.42}) and a two-weight equalizer (we use this channel and equalizer in figures throughout this thesis for comparison purposes). The labels "0", "1" and "2" are the locations of the MMSE weights for the corresponding sample delays. The minima are symmetrical about the origin due to the symmetric binary alphabet. Note that minima are close or coincident with the 0 and 1 sample delays for both algorithms. The CMA cost function has no minimum associated with the MMSE weights for a delay of 2 samples because the weights produce a negative SIR. DDA, however, has an additional very poor minimum that has the 2 sample delay within its region of convergence. Two element FIR channels with unit energy impulse response are determined by a single variable, i.e., h_m={sin(ϕ), cos(ϕ)}. We generated cost function plots for ϕ every five degrees and the results are always the same; a global minimum near the MMSE weights for the 0 or 2 delay (depending on whether h_0 or h_1 is maximum), and a poorer performing minimum near the MMSE weights for delay 1. DDA has a third very poor minimum except when the channel has low ISI (the minimum appears when ISI exceeds -12.7 dB). The existence of these bad minima is the reason DDA is considered unsuitable for training an equalizer

when the eye is not open. On the other hand, note that the MMSE points correspond exactly to the DDA cost function minima while the CMA minima are slightly offset from the MMSE points. Another factor, not apparent from the contour plots, is the convergence speed once the eye is open. DDA converges more quickly than CMA because the "bowl" around the minima is quadratic for DDA and quartic for CMA. A function proportional to x^4 has a much lower slope near zero than one proportional to x^2. These reasons are why DDA is preferred for final convergence.

For a more complex example, consider a channel with $h_m = \{1.16,\ 0.63,\ 0.08,\ 0.35,\ -0.70\}$, equalized with a 17 tap FIR filter. Figure 8a is a plot of the SIR of the MMSE weights versus delay. Note the very low SIR for long delays, and to a lesser extent, very short delays. It is important to note that no matter how many taps are used, there will be extreme delays that give poor performance yet may be stable points of Bussgang type algorithms. This property stems from the fact that the system of equations represented by Equation (18) is always underdetermined. Furthermore, the bad stable points are close to the good points (in Euclidean distance). To illustrate this property, Figure 8b shows the distance of the MMSE weights from the MMSE weights for the globally optimum delay of 8 samples.

We ran a simulation of the DDA algorithm for this example with the filter initialized to the MMSE weights for a delay of 0 samples. The initial MMSE weights produce an SIR of 5.4 dB and a BER of 0.031. The filter settled out to a stable point with an even worse BER of 0.043.

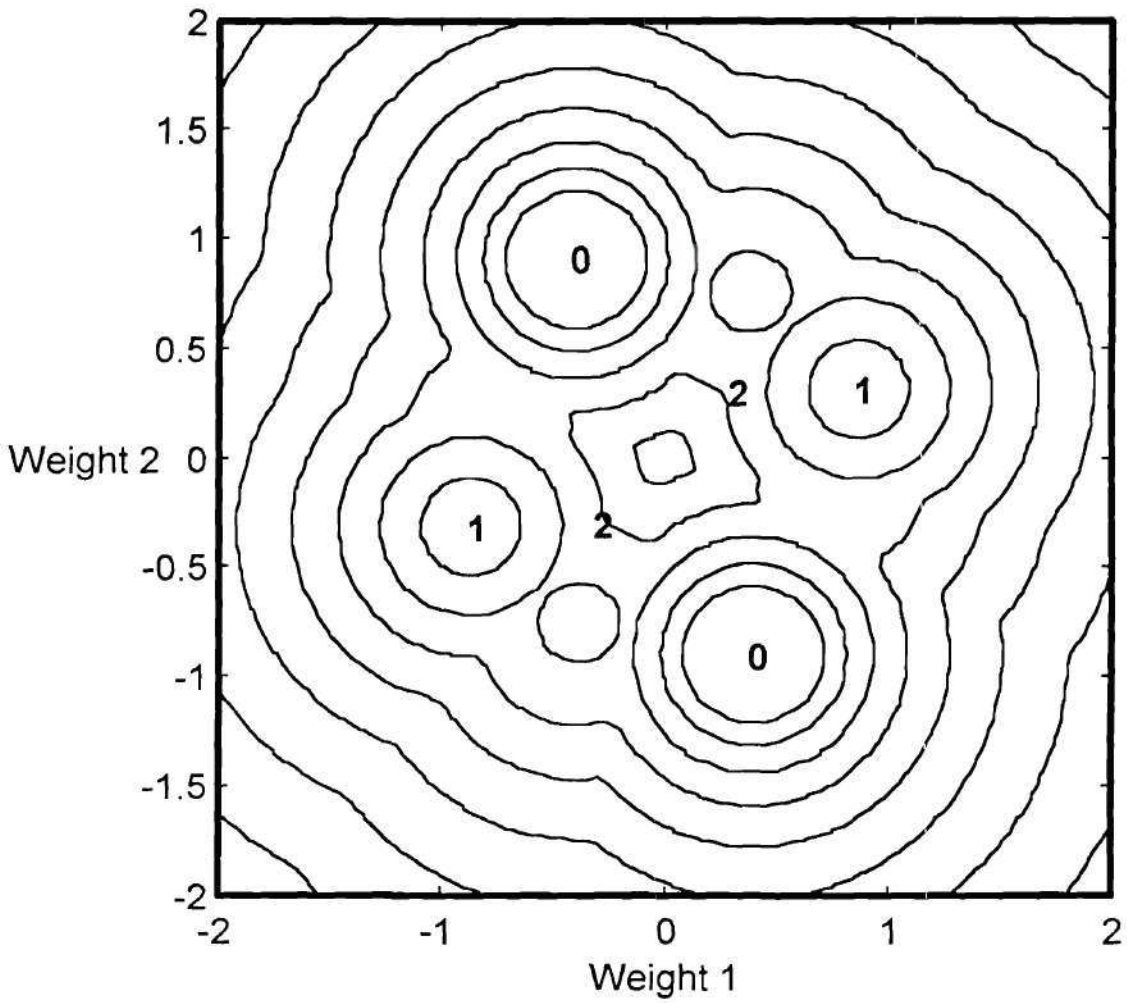


Figure 6: Contour plot of DDA cost function (h_m={1.0, 0.42}).

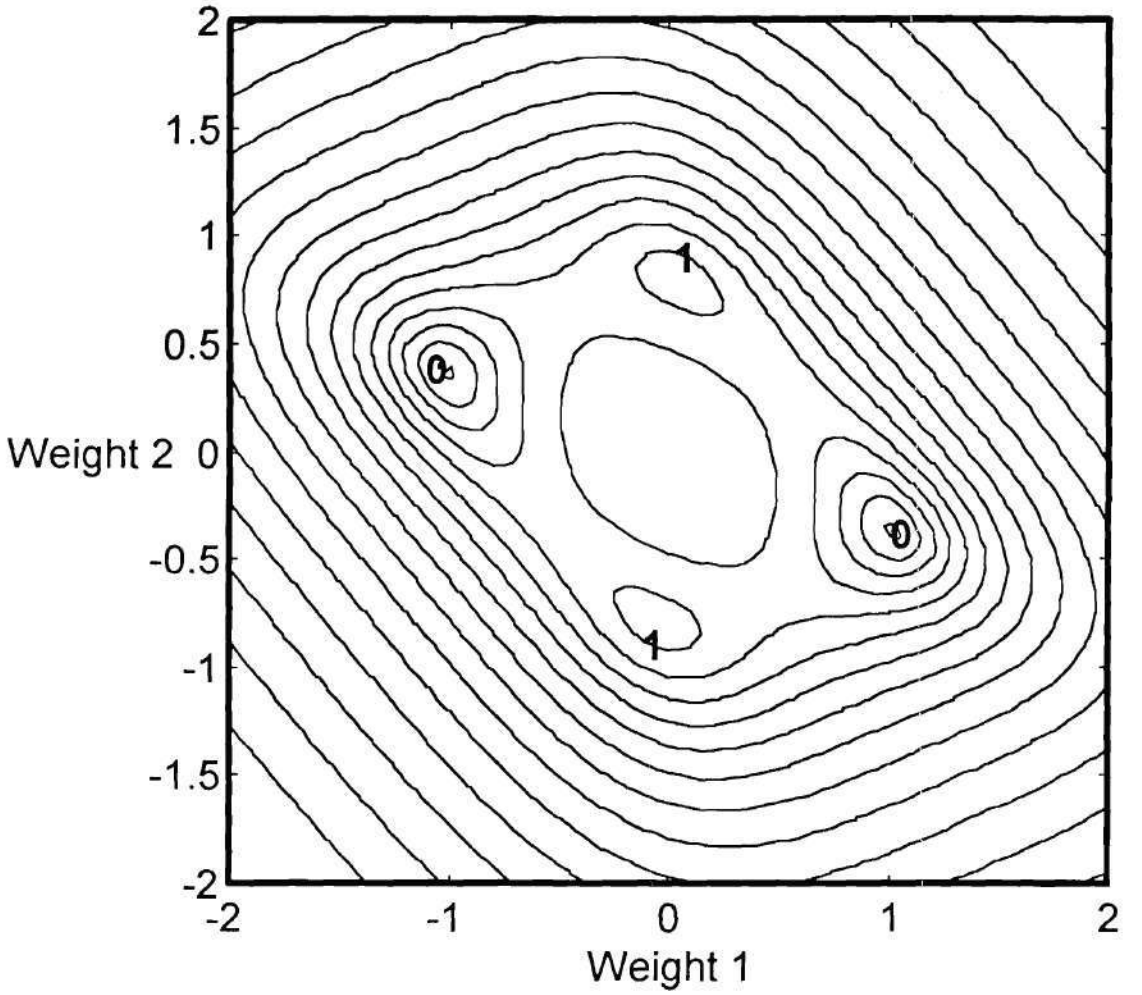


Figure 7: Contour plot of CMA cost function(h_m={1.0, 0.42}).

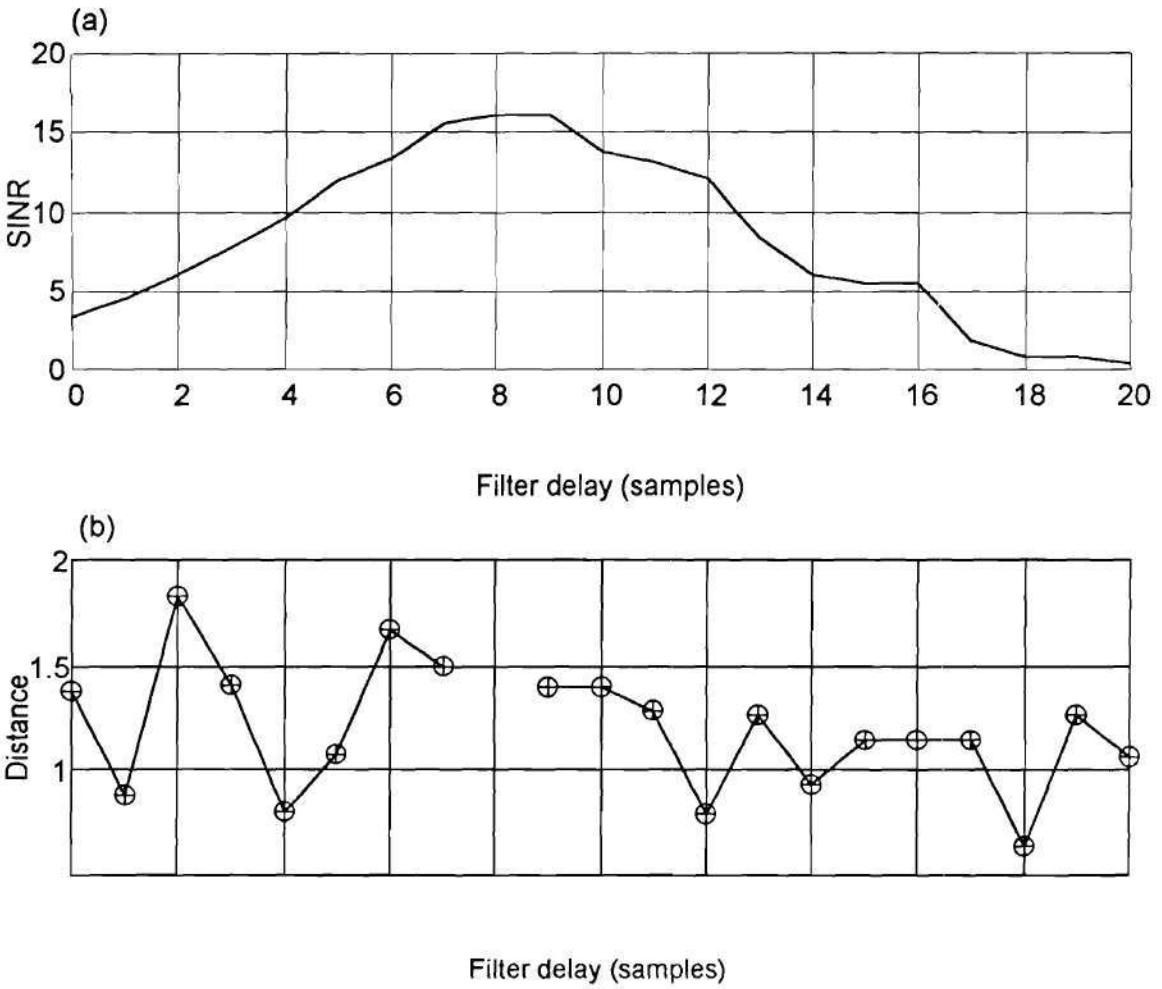


Figure 8: Example channel with 17 tap equalizer: a) Maximum SIR vs. delay, b) Euclidian distance between MMSE weights and weights for a delay of 8 samples.

3.1.1.1.Power Series Expansion

We have developed a novel analysis technique to show the importance of the MMSE weights. We examine the update term of the Bussgang algorithm with MMSE weights for a delay of d samples (the d^{th} column of $(\mathbf{H}^T)^{\#}$). We perform a power series expansion of the Bussgang error function, $e(y) = y - g(y)$, about the values of a_{m-d}, the data sequence delayed by d samples:

$$E\{\mathbf{x}_m e(y_m)\} = E\{\mathbf{H}\mathbf{a}_m e(a_{m-d})\}+E\{\mathbf{H}\mathbf{a}_m e'(a_{m-d})\Delta y_m\}+E\{\mathbf{H}\mathbf{a}_m e''(a_{m-d})\Delta y_m^2\}/2+ \ldots.(34)$$

where $\Delta y_m = y_m - a_{m-d} = \mathbf{I}_d^T(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{a}_m$ and $\mathbf{x}_m = \mathbf{H}\mathbf{a}_m$. Since the goal is to produce a pure delay, a_{m-d} represents a reference signal. We evaluated Equation (34) out to three terms (the Δy^2 term) and assumed a binary, {-1,1}, data sequence when taking expected values.

The first term is $E\{\mathbf{H}\mathbf{a}_m e(a_{m-d})\}=\mathbf{H}E\{\mathbf{a}_m e(a_{m-d})\}$. If $e(-1) = e(1)$ then $e(a_{m-d})$ is a constant and $\mathbf{H}E\{\mathbf{a}_m e(a_{m-d})\}= \mathbf{H}E\{\mathbf{a}_m\}e(a_{m-d})=\mathbf{0}$. Note that this term equals zero independently of $\mathbf{H}$. For reasons defined later, $e(y)$ must be an odd function. Therefore $e(-1) = e(1)$ only if both equal zero.

The linear term is:

$$E\{\mathbf{H}\mathbf{a}_m e'(a_{m-d})\mathbf{I}_d^T(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{a}_m\}=\mathbf{H}E\{e'(a_{m-d})\mathbf{a}_m\mathbf{a}_m^T\}(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{I}_d$$
$$=e'(a_m)E\{a_m^2\}\mathbf{H}(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{I}_d= e'(a_m)\,\mathbf{H}(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{I}_d=\mathbf{0}. \quad (35)$$

This equation equals zero if $e'(1) = e'(-1)$, which is always true if $e'(\bullet)$ is an even function (hence the requirement that $e(\bullet)$ be odd). In addition, $e'(1)$ and $e'(-1)$ must be negative for the MMSE point to be a local minimum instead of a maximum. The quantity $\mathbf{H}(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})$ is

always **0** because $(\mathbf{I}-\mathbf{H}^{\#}\mathbf{H}) \equiv \mathbf{P}_{\perp}$, the projection operator onto the nullspace of **H** (Proof: $\mathbf{H}(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I}) = \mathbf{H}\mathbf{H}^{\#}\mathbf{H}-\mathbf{H}\mathbf{I}=\mathbf{H}-\mathbf{H}=\mathbf{0}$).

The implications of the analysis of the first two terms of the expansion are as follows; suppose that the MMSE weights for a delay of d samples restrict the ISI (or crosstalk, depending on the application) such that the filter output is restricted to a small neighborhood about the reference sequence a_{m-d}. If $e(y)$ can be considered a linear function within that neighborhood, then the MMSE weights are a stable point of the blind algorithm. For these cases, there will also be a neighborhood about the MMSE weights (in weight space) that will converge to the MMSE weights. The particular nonlinear function, $e(y)$ determines the size of the "linear neighborhood".

For example, DDA has $e(y) = y\text{-sgn}(y)$ and is linear everywhere except for a discontinuity at the origin. It has the largest possible linear neighborhood about the reference sequence so in this sense it is optimum; however, the discontinuity makes analysis difficult when the ISI exceeds the neighborhood. Clearly, the MMSE weights will be stable points for DDA if they open the eye. DDA is preferred for final convergence because of this property; if the eye is opened it behaves as if a training sequence is present and converges exactly to the MMSE point.

For the third term first recognize that $e''(y)$ is odd so $e''(a_{m-d})=e''(1)a_{m-d}$. Remembering this fact, term 3 is:

$$\mathbf{t}=\frac{e''(1)}{2}\mathbf{H}E\{\mathbf{a}_m\mathbf{I}_d^T(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\,\mathbf{a}_m\mathbf{a}_m^T(\mathbf{H}^{\#}\mathbf{H}-\mathbf{I})\mathbf{I}_d a_{m-d}\} = \frac{e''(1)}{2}\mathbf{H}E\{\mathbf{a}_m\mathbf{I}_d^T\mathbf{P}_{\perp}\mathbf{a}_m\mathbf{a}_m^T\mathbf{P}_{\perp}\mathbf{I}_d a_{m-d}\}. \quad (36)$$

Evaluating term 3 element by element then reassembling the vector gives (see Appendix A):

$$\mathbf{t} = \frac{e''(1)}{2}(p_{dd} - 2p_{dd}^2)\mathbf{H}_d = \frac{e''(1)}{2}\left(\frac{1}{\mathrm{SIR}_d+1} + \frac{2}{(\mathrm{SIR}_d+1)^2}\right)\mathbf{H}_d \tag{37}$$

where p_{dd} is the d^{th} diagonal element of $\mathbf{P}_\perp$ and SIR_d is the SIR produced by the MMSE weights for a delay of d samples. If SIR_d is high (say, >20 dB) Equation (37) may be approximated as

$$\mathbf{t} \approx \frac{e''(1)}{2\mathrm{SIR}_d}\mathbf{H}_d . \tag{38}$$

The third term in the expansion creates a bias term that is in the direction of $\mathbf{H}_d$, proportional to the second derivative of $e(y)$ and inversely proportional to the SIR produced by the MMSE weights. Therefore, if the crosstalk distribution exceeds the linear neighborhood, a bias term arises that drives the weights away from the MMSE weights. If the bias from the third and higher terms is sufficiently small, then the algorithm will settle to a point that still can be considered an approximation to the pure delay, i.e., the MMSE point is within the region of convergence of a stable point. The stable point will have even more MSE and probably increase the BER. Some of the stable points can have high residual SIR. This is the cause of at least some of the observed misconvergences.

To summarize: if $e(\bullet)$ is odd and passes through zero at ± 1 with a negative slope it will have stable points near the MMSE points as long as $e(\bullet)$ is reasonably smooth (i.e., the higher order terms of the expansion can be neglected). In fact, any odd function that crosses zero with a negative-slope will be stable near scalar multiples of the MMSE points. If the zero-crossings occur at $\pm c_0$ then the stable points will be near the columns of $c_0(\mathbf{H}^T)^{\#}$.

3.1.1.2.Monte-Carlo Simulation

We now describe a Monte-Carlo simulation of 3074 channels chosen at random. We wanted to test the hypothesis that initialization with low SIR MMSE weights leads to misconvergence. Each channel has a real valued, 4-element impulse response, such that one of the four elements (randomly selected) is set to one and the other three are chosen from a Gaussian distribution with standard deviation of ¼. Each channel is equalized with a 6 tap filter using both the CMA and Decision-Directed cost functions. We calculated the MMSE weights and the SIR for each of the 9 possible delay values, then selected a delay corresponding to a poor SIR. For CMA we chose the smallest positive SIR value, assuming that the negative SIR delay values could not be stable. We have observed that minima with negative SIR often exist for the DDA (this correlates with the fact that the DDA is not used in practice to initialize equalizers). Therefore, for DDA we tried both the smallest positive and largest negative SIR values and then selected the one that resulted in the worst performance. The algorithms were initialized with the weights and allowed to converge. The weights were updated using the actual gradient of the cost functions instead of the instantaneous estimates used in stochastic gradient algorithms to get a truer estimate of the characteristics of the underlying cost functions.

For *all* channels and for both cost functions the weights settled to nearby points that had an SIR equal to or slightly worse than the initial points. In other words, the MMSE points were *always* in the region of convergence of a low SIR stable point as we predicted. An interesting example is trial number 2924, h_m={-0.39, -0.63, 1.00, -0.46}, the bad MMSE weights produce an SIR of 1.7 dB and the CMA cost function converges to a point with a SIR of -0.67 dB. So local minima exist in the CMA cost function even for very low SIR points. Figure 9 and Figure 10 show the SIR produced by the bad points versus the best possible SIR that could be produced by any weights for DDA and CMA

respectively. The figure displays both smoothed and unsmoothed data. The data was smoothed with a 50 point Kaiser window with β=5. Note that the SIR produced by the bad points is much less than the global maximum SIR. DDA gives even worse results than CMA because it often has worse minima that produce a negative SIR.

3.2. Variance Constraint Cost Functions

3.2.1. Previous Work

Since the idea of blind equalization was introduced, authors have proposed techniques for avoiding misconvergence. [Sato '75] and [Godard '80] initialized their algorithms with a single nonzero tap; this approach is sensitive to the overall channel gain. Since the location of the minima is determined by the channel matrix **H**, all else being equal, their distances from the origin are inversely proportional to the overall channel gain. For example, assume that $\mathbf{W}_0$ is a minimum for a given cost function and a channel **H**. If **H** is multiplied by a constant c then the minimum move to a new location $\mathbf{W}_0/c$. If the gain is low (e.g., during a deep fade) the minima are far from the origin and the local maximum at the origin is much larger (refer to Figure 6 and Figure 7). In that case a fixed initialization will seem to be near the origin and near the local maximum. The regions of convergence for all the minima converge at the origin so the vagaries of the random data may push the weights into a bad minimum. On the other hand, if the channel has a high gain, the minima will have a small magnitude and the fixed initial value will appear far outside the collection of minima. Again, the actual data sequence could steer the weights to a bad minimum.

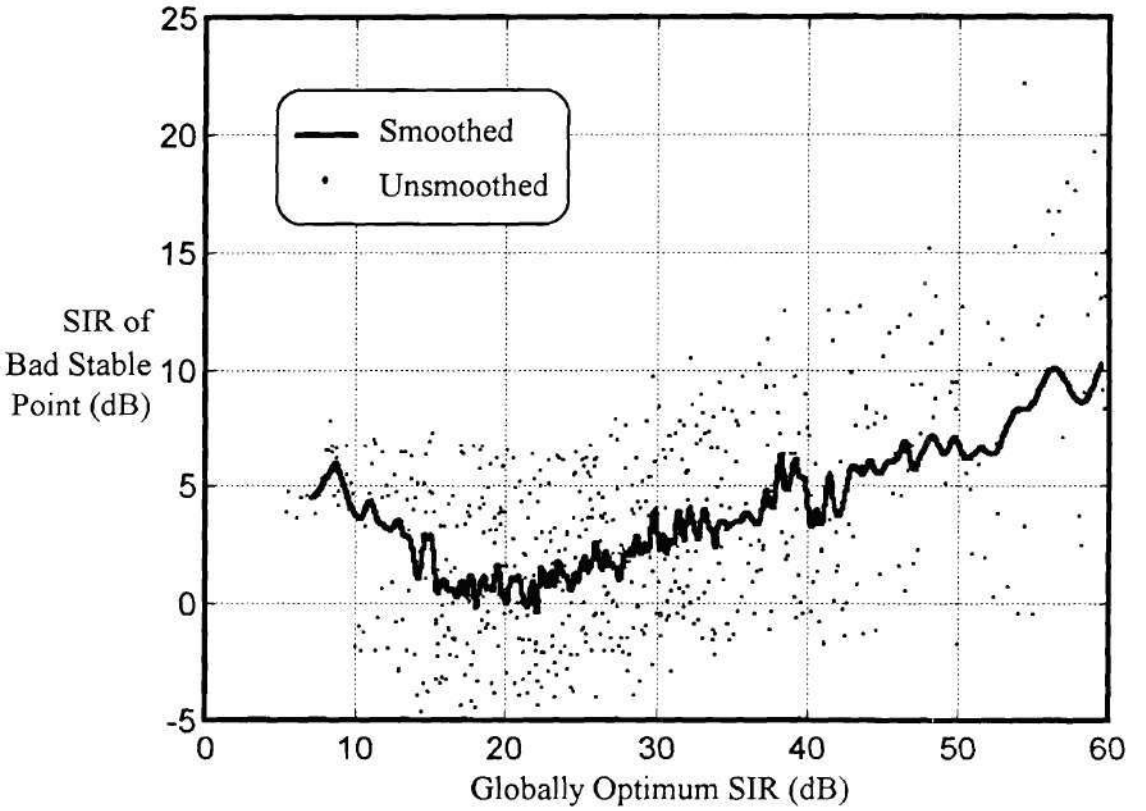


Figure 9: SIR of bad stable points versus the global maximum SIR for DDA, result of 3074 randomly generated channels, data smoothed by 50 point Kaiser window with $\beta = 5$. Every fifth unsmoothed data point is plotted.

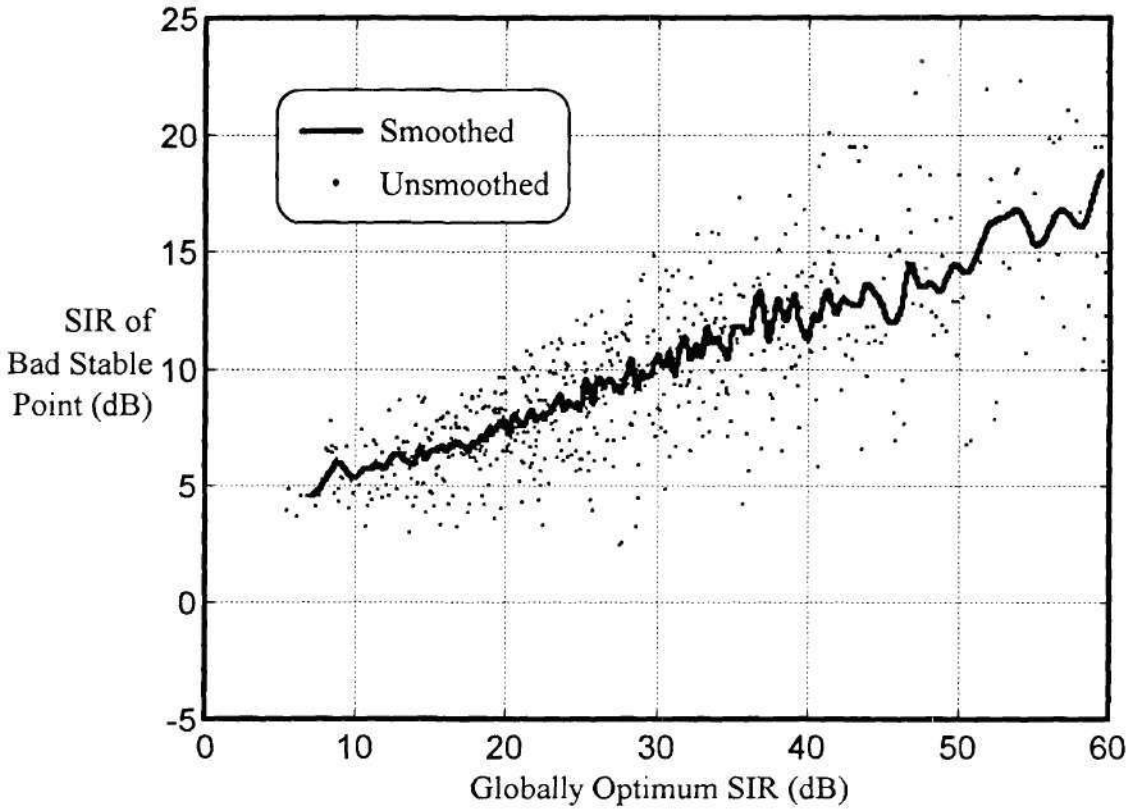


Figure 10: SIR of bad stable points versus the global maximum SIR for CMA, result of 3074 randomly generated channels, data smoothed by 50 point Kaiser window with $\beta = 5$. Every fifth unsmoothed data point is plotted.

We studied the deep fade case with a Monte-Carlo simulation. The randomly produced channels had the same statistics as the previously described Monte-Carlo simulation, except that the equalizer had 7 taps instead of 6. After generating 2050 channels, we rejected any channel that did not allow at least 20 dB SIR. 1669 channels remained. A 20 dB fade was simulated by multiplying **H** by a factor of 0.01. DDA and CMA were simulated on each channel. We initialized the weights to $\mathbf{I}_3$ and updated them using instantaneous estimates from random binary data. We declared misconvergence if the algorithm converged to a stable point that yielded an SIR of less than 10 dB. Out of the 1288 trials, DDA failed 510 times (31 %) and CMA failed 103 times (6 %).

[Foschini '85] used a tap centering procedure that monitors the center of gravity of the tap weights and recenters the weights periodically. This technique seems to work well, but theoretical underpinnings are lacking [Ding '91A]. We will not consider it any further for equalization in this thesis, although we consider a similar technique we call "peak forcing" for use with WDM crosstalk cancellation.

Recent work has been done with linear constraints (also called anchors), where the cost function is minimized under the constraint $\mathbf{W}^T\mathbf{C} = 1$. The constraint restricts the weights to an affine space orthogonal to **C**. Such a scheme can be implemented using the GSC architecture discussed in Chapter 2 and is the basis for the robust performance seen in the CDD crosstalk cancellation algorithm considered in the next chapter. The CDD algorithm assumes a symbol alphabet with zero as one of the symbols. The presence of a zero allows a simple method for eliminating the desired signal from the lower branch of the GSC.

[Kamel '94] claimed that an equalizer will exhibit global convergence properties when the standard CMA cost function is used along with a constraint, as long as the channel has at least a minimum gain. To see why return to Figure 7 and look at the 2-

dimensional cost functions. Assume a simple constraint that restricts the weights to a vertical line, $w_1=a$. If a is large the constrained cost function has a single minimum where the constraint line intersects the radial line connecting the origin with the global minimum. Any point along the radial line will have maximum SIR because SIR is independent of weight magnitude. If a is smaller than one the constrained cost function will lose this desirable property and will have two minima near the minima associated with the delay-1 MMSE weights. Therefore, performance of linearly constrained DDA and CMA is sensitive to channel gain. Figure 11 shows cuts through the cost function of Figure 7 for various values of a.

[Verdu '93] showed that global convergence can be achieved if the first tap is held to one and the standard CMA cost function is replaced by a minimum energy cost function $J = E\{y^2\}$. If the channel is non-minimum, phase this technique requires an additional allpass filter. The allpass must be made up of a subset of the zeros of the adaptive filter (and their reflections outside the unit circle). Verdu proposed that the allpass section be determined by first solving for the zeros in the adaptive filter and then switching in allpass sections containing a real zero or a conjugate pair. The overall filter output must be checked for each combination to see if the equalizer output has the proper distribution.

If the filter is large, the number of possibilities for the allpass section can become prohibitively large. An N tap filter has $2^{N/2}$ possibilities (assuming the best case scenario of all complex zeros). For example, a 20 tap filter has 1024 possibilities. Also, many techniques already exist for using second order statistics to determine the magnitude response of a channel, which is really all that the "blind" portion of the filter is doing.

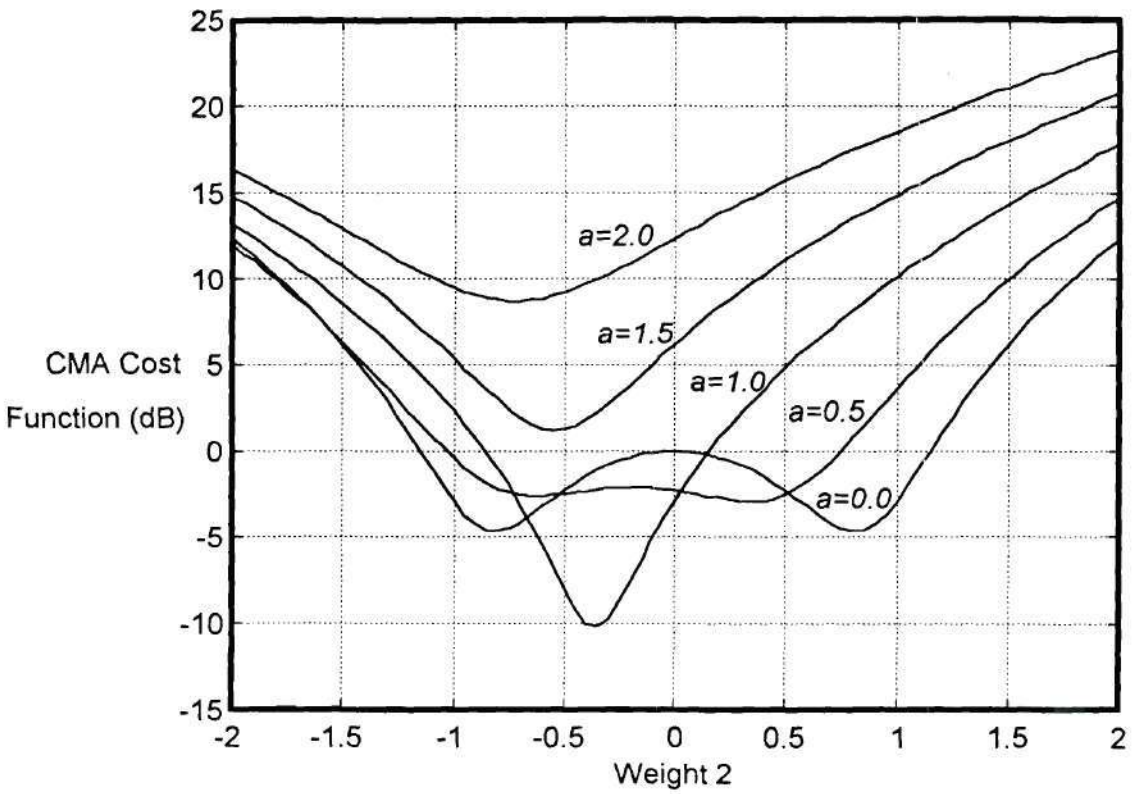


Figure 11: One dimensional cuts through the CMA cost function of Figure 7. Weight one is held constant at 0.0, 0.5, 1.0, 1.5, 2.0. Note failure of function to have a single minimum for the 0.0 and 0.5 cases (h_m={1.0, 0.42}).

The work was extended in [Vembu '94] to show that a cost function of $J = \sum|t_i|$, (the l_1 norm of the product $\mathbf{W}^T\mathbf{H}$) and nearly any constraint will produce a global minimum. Figure 12 shows this cost function for the example channel and equalizer (h_m={1.0, 0.42}, two weight equalizer). Note that, unless the constraint is parallel to one of the linear sections, only one minimum exists. Vembu shows that the l_∞-norm of the output y_m is equivalent to the cost function. This quantity is approximated by calculating $E\{|y|^p\}$ with p large (Vembu used 8 to 14 used in the paper). High exponentiation means that the stability of an LMS-type update equation based on this cost function will be extremely sensitive to step size and/or channel gain.

3.2.2. Variance Constraint Algorithms

We propose a new class of Bussgang-type equalizers, called Variance Constraint (VC), that have cost functions designed to operate with linear constraints in the weight space. The constrained VC cost functions have an advantage over standard Bussgang cost functions because the constraint relocates the low-SIR convergence points farther away than those for the Bussgang algorithms that cannot use constraints. Specifically, If $\mathbf{W}_d$ is the d^{th} column of $(\mathbf{H}^T)^{\#}$, then the VC minima will be near $\mathbf{W}_d/(\mathbf{C}^T\mathbf{W}_d)$, i.e., the unconstrained minima are scaled until they intersect the affine space and satisfy the constraint. If a bad minimum is nearly orthogonal to $\mathbf{C}$ the denominator is small and the point is pushed far away. Figure 13 illustrates the effect for a simple 2-delay, 2-weight system. Figure 13a) shows the columns of the $\mathbf{H}$ matrix and the MMSE weights. Figure 13b) shows the case where a constraint of $[1,0]^T$ is used. Note that the weights for delay one are pushed away from the weights for delay zero. Figure 13c) shows the optimum case where the constraint equals $\mathbf{H}_0$. In this case the line to which the weights are constrained is parallel to MMSE weights for channel 2 so the channel 2 weights are pushed to infinity (neglecting noise).

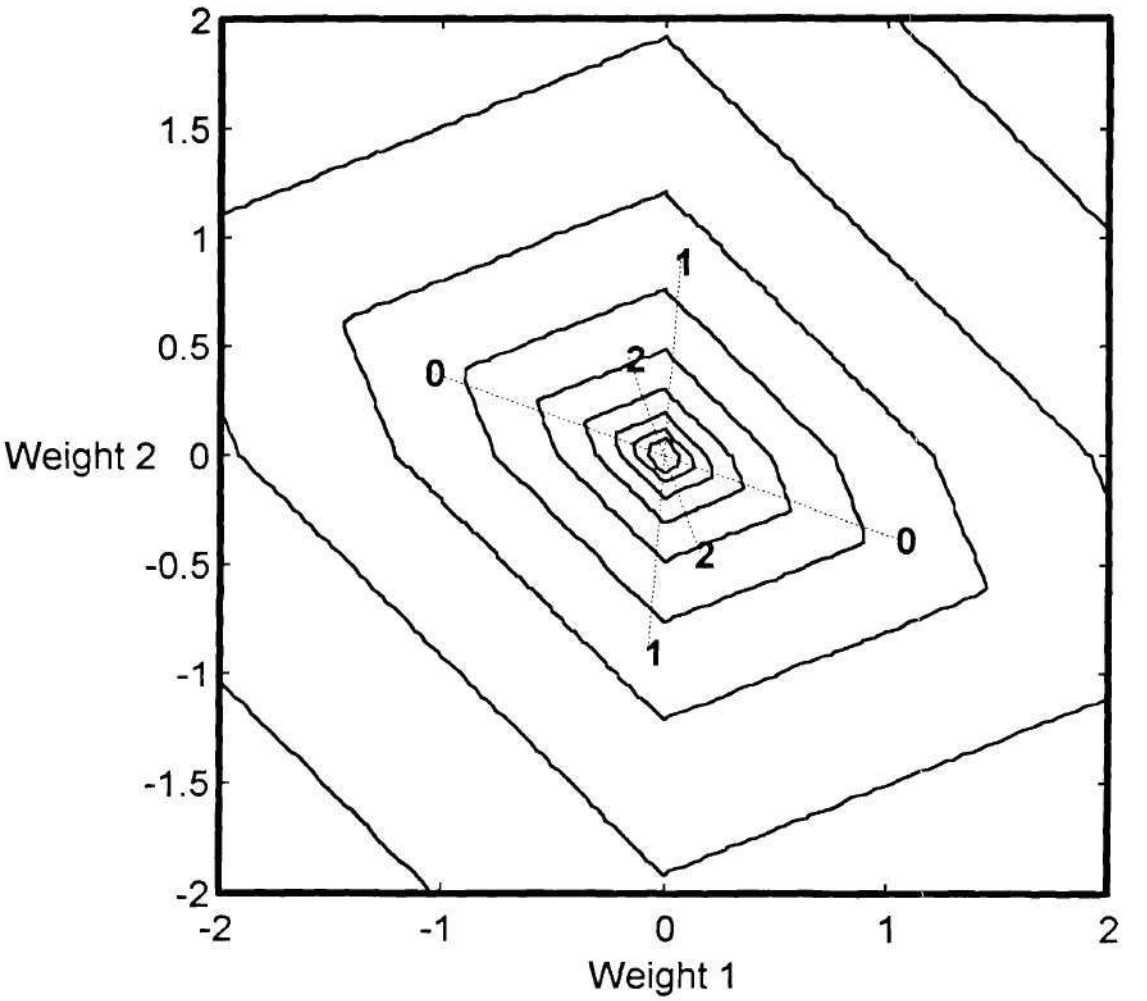


Figure 12: Contour plot of l_1 cost function proposed in [Vembu '94] (h_m={1.0, 0.42}).

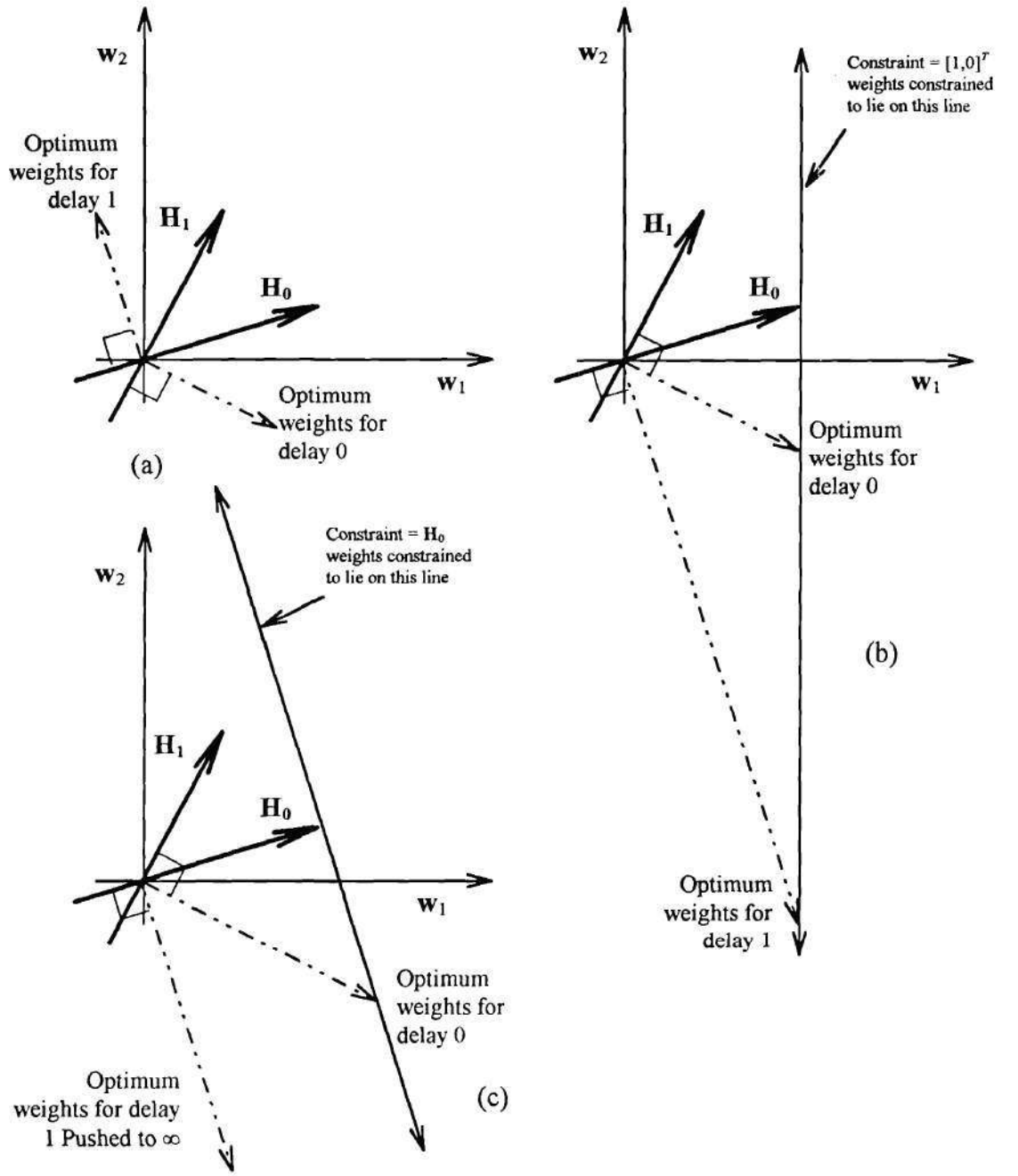


Figure 13: Effect of constraint on misconvergence: a) no constraint, b) constraint of $w_1 = 1$, c) constraint of $\mathbf{H}_0$.

Figure 14 shows this effect for the channel h_m = {-0.17, 0.30, 1.00, -0.24, -0.15}, and a 13 tap equalizer where the seventh tap was anchored to one. For each MMSE point we plotted its Euclidean distance from the anchor $\mathbf{I}_6$ versus the SIR produced by the weights for both the unconstrained and constrained cases. Note that the low SIR MMSE points are far away from $\mathbf{I}_6$ when a constraint is used. The short vertical line segment at the end of the plot is because the MMSE weights closest to $\mathbf{I}_6$ produces a slightly lower SIR than the maximum SIR MMSE weights.

We modified the Godard cost functions to create the VC cost functions. Recall from Chapter 2 that the Godard dispersion function of order p is $J_G^p = E\left\{\left(|y|^p - c\right)^2\right\}$, where p=1,2,... and c is a constant. The corresponding VC_p cost function, is:

$$J_{VC}^p = E\left\{\left(|y|^p - E\left\{|y|^p\right\}\right)^2\right\} = \sigma_{|y|^p}^2, \tag{39}$$

i.e., the variance of the output magnitude raised to the p^{th} power. The only difference is that we replaced the constant c by $E\left\{|y|^p\right\}$; thus, the overall channel-equalizer gain is constantly being estimated and incorporated into the cost function. In practice, we require time averages to calculate expected values, therefore, memory is incorporated into the cost functions and update equations. The inclusion of memory, plus the requirement for a constraint means that the VC algorithms are not Bussgang algorithms. We will restrict our study to p equal one or two.

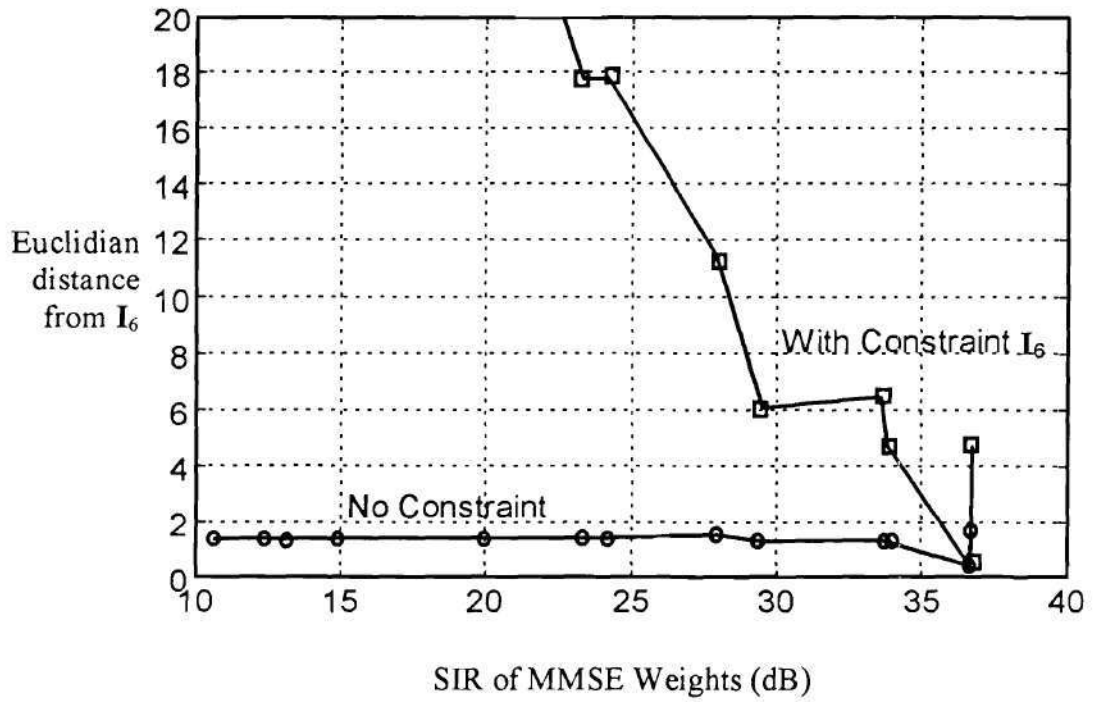


Figure 14: Euclidian distance from $\mathbf{I}_6$ vs. SIR for MMSE weights of an example channel, $h_m = \{-0.17, 0.30, 1.00, -0.24, -0.15\}$, with a 13 tap equalizer.

For p=1, the algorithm is denoted VC_1, the cost function is $J_{VC}^1 = E\left\{\left(|y| - E\{|y|\}\right)^2\right\}$, and the gradient is $\nabla_{\mathbf{W}} J_{VC}^1 = E\left\{\mathbf{x}\left(y - E\{|y|\}\,\mathrm{sgn}(y)\right)\right\}$, so the update equation is:

$$\mathbf{W}_{m+1} = \mathbf{W}_m - \mu\mathbf{x}_m(y - E\{|y|\}\mathrm{sgn}(y)). \tag{40}$$

$E\{|y|\}$ must be estimated using some sort of filter to get a time average, which would tend to slow down convergence.

For p equal 2, denoted VC_2, the cost function is $J_{VC}^2 = E\left\{\left(y^2 - E\left\{y^2\right\}\right)^2\right\}$, the gradient is $\nabla_{\mathbf{W}} J_{VC}^2 = E\left\{\mathbf{x}\left(y^3 - E\left\{y^2\right\}y\right)\right\}$ so the update equation is:

$$\mathbf{W}_{m+1} = \mathbf{W}_m - \mu\mathbf{x}_m(y^3 - E\{y^2\}y). \tag{41}$$

$E\left\{y^2\right\}$ can be formed using the relationship $E\left\{y^2\right\} = \mathbf{W}^T\mathbf{R}_{xx}\mathbf{W}$, $\mathbf{R}_{xx}$ can be estimated independently of the weights so the step size is unrestricted. If the channel is stationary, $\mathbf{R}_{xx}$ need only be estimated once and then it can be used throughout the session; otherwise, it maybe updated in parallel with the weights. Neither method will slow convergence.

Figure 15 and Figure 16 shows the unconstrained VC cost functions for the example channel; Figure 15 is VC_1 and Figure 16 is VC_2. Note the single global minima at the origin and the lack of local minima at the MMSE points. The minima have been replaced by "valleys" or "troughs"sloping to the origin. The centers of the troughs run along radial lines that pass through the origin and the minima of the corresponding Godard cost function. The points along a trough are scale multiples of a Godard minimum and

hence have the same SIR. We have already shown that the Godard minima are nearly coincident with the MMSE points if they produce a high SIR. So in that case, any point lying along the center of the trough will also produce a high SIR.

3.2.2.1.Constraint Selection

The VC cost functions require constraints; without them the weights will go to the global minimum at the all zero vector. The minima of the constrained cost function occur when the constrained affine subspace intersects the troughs. The choice of constraint is problematic. For example, Examination of Figure 16 shows that a constraint of w_1=1 settles to the maximum SIR weights for delay 0 and a constraint of w_2=1 settles out to the poorer weights for delay 1. So the desire is to select a constraint that makes a small angle with "good" MMSE points and has a large angle with bad points. The best constraint for both cost functions is one of the center columns of **H** because then both algorithms will return the MMSE weights for the corresponding delay and the SIR is maximized. Of course, such knowledge is unlikely to be available and would be equivalent to identification of the channel. A good strategy is setting one of the center taps to 1 and the rest to 0. The strategy works because the bad points are related to extreme delays so they will have the hsignificant weight values bunched to one end of **W**. Therefore, if enough taps exist, the bad-delay minima will be nearly orthogonal to **C** and be pushed far away.

If the channel is very stable and is not expected to change much from session to session, the optimum constraint can be derived from the current adaptive weights. The quantity $\mathbf{R}_{xx}\mathbf{W}$ is approximately equal to $\mathbf{H}_d$ if the SINR is high:

$$\mathbf{R}_{xx}\mathbf{W} = \mathbf{H}\mathbf{H}^T\mathbf{W} = \mathbf{H}\mathbf{T} \approx \mathbf{H}\mathbf{I}_d = \mathbf{H}_d. \tag{42}$$

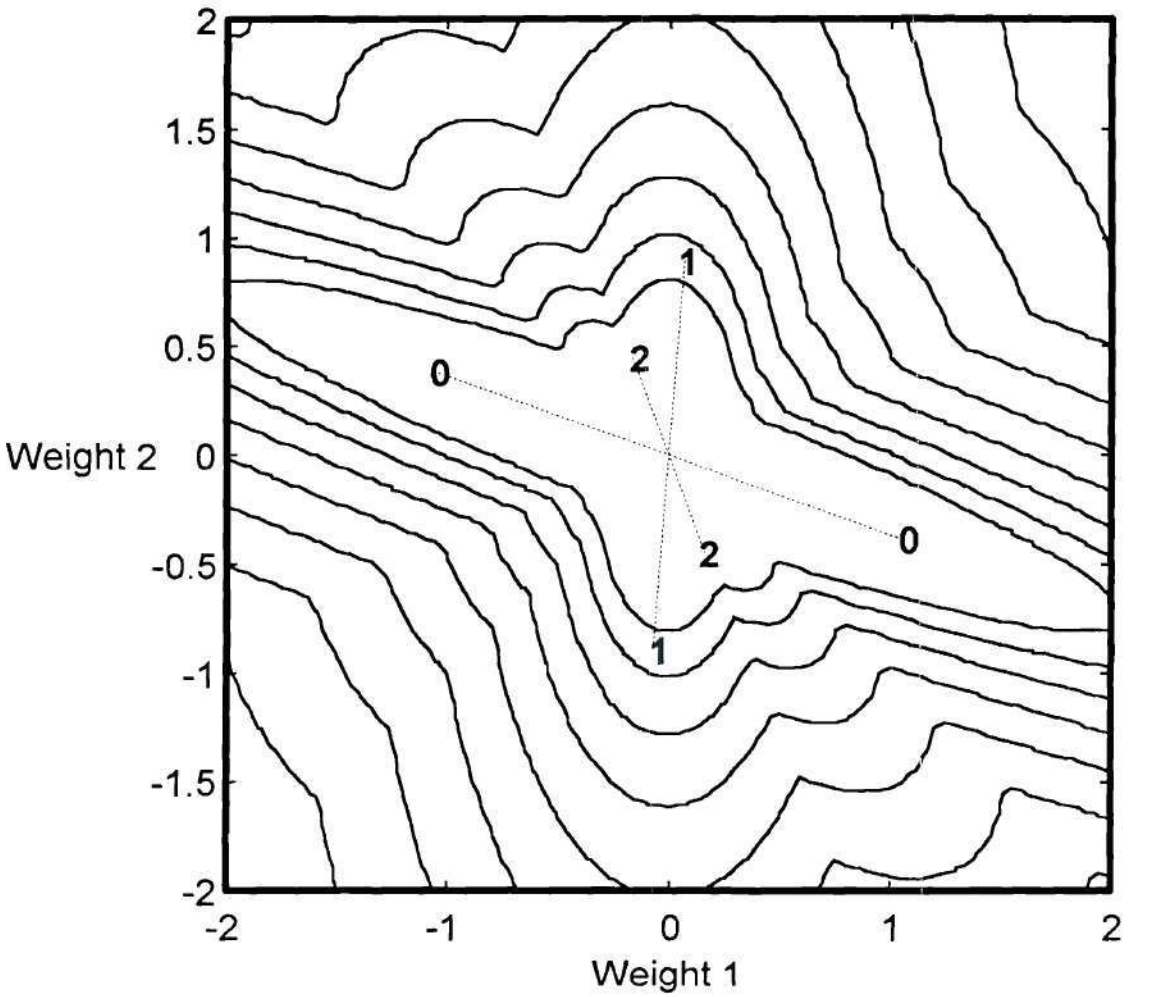


Figure 15: Contour plot of VC_1 cost function (h_m={1.0, 0.42}).

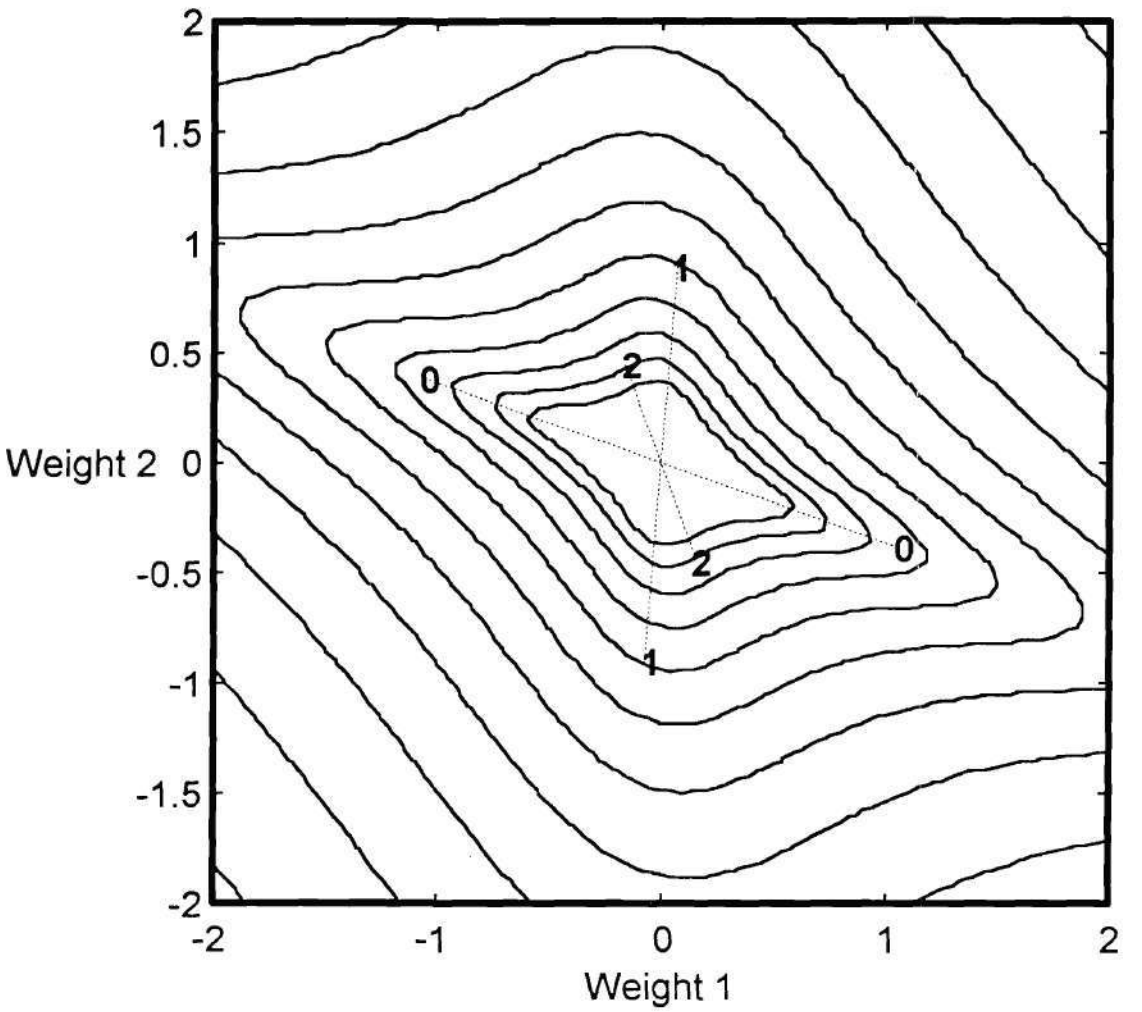


Figure 16: Contour plot of VC_2 cost function (h_m={1.0, 0.42}).

Figure 17 shows both the VC_1 and VC_2 cost functions constrained to w_1=1 for the example channel (we scaled the Magnitude of the VC_1 curve by 20 for plotting purposes). Note the presence of a single global minimum, which coincides with the location of the maximum SIR weights. This is because, for this particular example, h_0 is the largest magnitude value of the impulse response h_m. A constraint of $\mathbf{I}_0$ is always a multiple of $\mathbf{H}_0$ (see Equation (19)), therefore, a constraint of $\mathbf{I}_0$ is the optimum constraint for selecting a delay of zero samples because it creates the situation shown in Figure 13c. If $|h_0|$ is maximum then the MMSE weights for a delay of zero samples produces a high SIR. Unfortunately, if $|h_0|$ is not the maximum the MMSE weights for zero delay is very poor so this constraint is not a good choice for unknown channels.

If we consider the dynamic behavior of the adaptive algorithm, then we find another effect that lessens the chance of misconvergence. For a fixed step size, μ, the high SIR minima may be stable while the bad minima are unstable. The weights for the bad minima will have a large magnitude (assuming a well-chosen constraint). The VC cost functions all monotonically increase with magnitude, more specifically $J_{VC}^{p}(c\mathbf{W}) = c^{2p} J_{VC}^{p}(\mathbf{W})$, hence the gradient also increases with weight magnitude. Therefore, the update term in Equation (20) becomes very large near the bad minima. Taking large noisy steps, the algorithm can step out of the region of convergence of the bad minima. So a step size that is stable for a good minima of small magnitude may very well be unstable for a large magnitude minima. Therefore, even if the point is stable in theory, it may not be stable for a realizable stochastic algorithm. We have observed this effect in simulations. Even when the algorithm was seeded with a candidate bad point the weights "kicked" out from the region of convergence and settled into a better performing point with smaller magnitude. Contrast this to the Godard algorithms, which do not have this property because the weight vectors that produce minima all have comparable magnitudes.

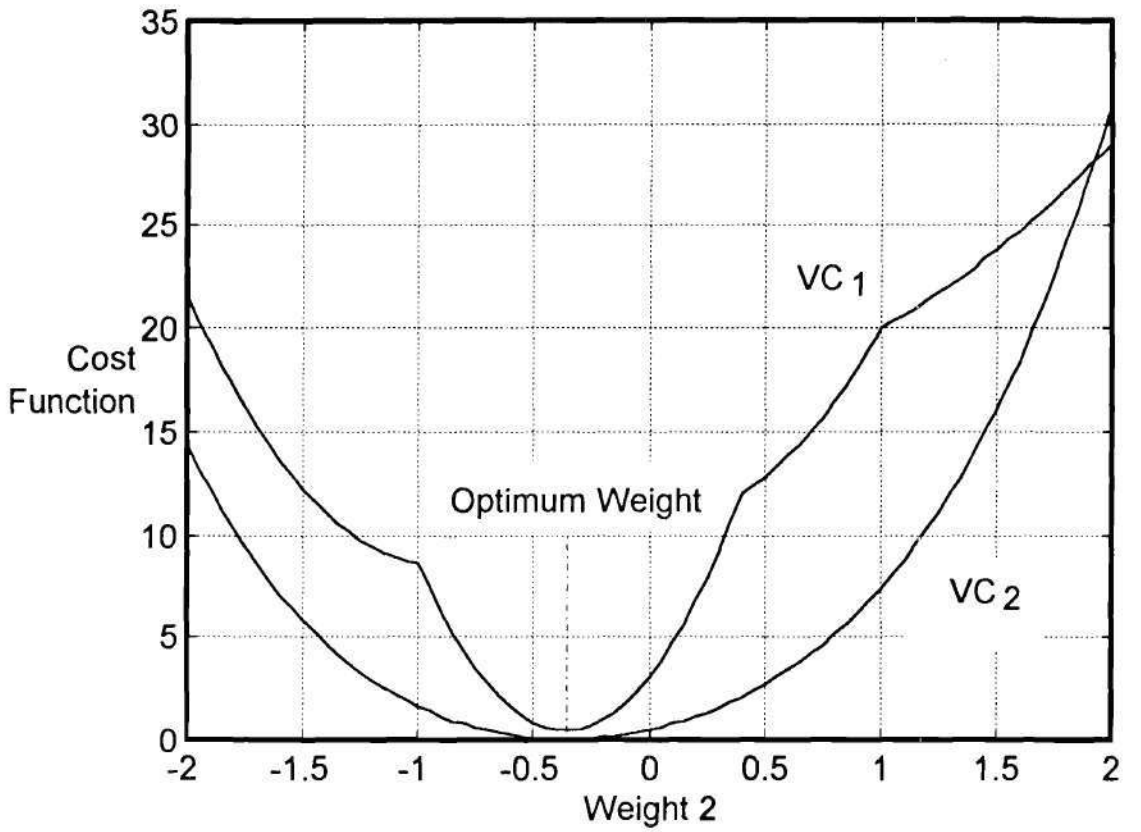


Figure 17: Cost functions of VC_1 and VC_2 from Figure 15 and Figure 16 constrained to w_1=1. The magnitude of the VC_1 curve is scaled by 20 for plotting purposes (h_m={1.0, 0.42}).

3.2.2.2.Monte-Carlo Simulation of VC Algorithms

We performed a Monte-Carlo simulation of both the VC_1 and VC_2 algorithms. The channels were the same channels generated for the deep fade Monte-Carlo simulation described earlier in this chapter. We rejected any channel that could not produce at least 15 dB SIR for some weight vector, leaving 1987 channels. The constraint in all cases was $\mathbf{I}_3$, which also served as the initial weight setting. We updated the weights using instantaneous estimates from random binary data. For the VC_1 algorithm we estimated $E\{|y|\}$ by averaging 20 filter outputs between each weight update. For the VC_2 algorithm we assumed that $\mathbf{R}_{XX}$ was known and used $E\{y^2\}=\mathbf{W}^T\mathbf{R}_{XX}\mathbf{W}$. The simulations lasted 350 iterations. We ran any cases that had yet to converge additional iterations until convergence was reached. This happened in less than 5 % of the cases. After convergence, we compared the weights to the MMSE weights which were closest to the constraint $\mathbf{I}_3$. For all channels and for both algorithms the weights converged to the MMSE weights that were closest to the constraint vector as hoped.

CHAPTER 4.

ALGORITHMS FOR WDM CROSSTALK CANCELLATION

We have developed four algorithms for WDM crosstalk cancellation. Two come from the realm of blind equalization, DDA and CMA augmented with a peak forcing algorithm. Two have aspects that make them new. The Constrained Decision Directed algorithm is designed to work with on-off-keying (OOK) signals. We have shown that CDD is closely related to the VC_1 algorithm described in the previous chapter. The pilot tone algorithm is not a blind algorithm; however, data transmission does not have to be interrupted to set the weights. Thus, it shares an important attribute with blind algorithms.

4.1. Algorithms Derived From Blind Equalization

Existing blind equalization algorithms may be applicable to the crosstalk cancellation problem.

4.1.1. Applicability of Blind Equalization Algorithms to Crosstalk Cancellation

The Bussgang algorithms hold the most promise for crosstalk cancellation. They are simple and by filtering out the d.c. component of the detector currents[1] all the Bussgang techniques can be applied to crosstalk cancellation. Special weight management techniques must be developed to ensure that the filter always selects the desired channel, because in the crosstalk case, variable "delay" means selecting the wrong channel.

---

[1]Identical notch filters may be added at the output of each photo-detector. However; if the time constant of the notch filter is much less than the convergence time of the equalizer, a single notch filter may be placed at the output of the equalizer.

Of all the Bussgang algorithms described in Chapter 2, only two apply to ±1 data: decision directed and constant modulus. The Sato algorithm, the decision directed algorithm, and hence, the combined Sato/decision-directed algorithms are all identical for the case of binary data. The Bellini algorithm is specifically designed for multilevel PAM (the development relies on approximating the data symbol distribution as uniform).

Explicit HOS methods require on the order of N^3 operations each iteration [Proakis '91], and are not suitable for the crosstalk application, where a premium is on low complexity algorithms. We envision the adaptive algorithm integrated on the same chip as the detector array and implemented partially with analog electronics [Hirotsu '93]. Thus, anything as complex as the explicit HOS would be inappropriate.

4.1.2.Relationship of Blind Equalization and Crosstalk Cancellation

We now consider the Decision Directed and Constant Modulus algorithms as algorithms for crosstalk cancellation in WDM receivers.

We now compare the matrix representation of the equalizer output and the crosstalk canceller. For the equalizer, $y_m=\mathbf{W}^T(\mathbf{H}\mathbf{a}_m+\mathbf{n}_m)$ and for the crosstalk canceller $y=\mathbf{W}^T(\mathbf{GS}+\mathbf{n})$. Some remarks follow.

The symbols used in the fiber optic networks are {0,1} and are not zero mean, which violates an assumption of every blind equalization paper in the references. Passing the canceller output through a filter to remove the d.c. component will eliminate this distinction (because the channel and filter are assumed LTI, removing the mean from the output is equivalent to removing the mean from all of the inputs).

For crosstalk applications, it is possible for the number of weights to equal or exceed the number of channels. only a finite number of channels exist, compared to an infinite number of possible delays of the received signal. So **G** can be overdetermined, underdetermined or uniquely determined depending on the number of channels and the

number of detectors. Unlike the equalization problem where **H** is underdetermined, the crosstalk can be completely cancelled if **G** is uniquely determined or overdetermined. If the number of detectors match or exceed the number of channels this condition should be met in a well designed receiver. For this case the CMA algorithm is guaranteed to converge to plus or minus a pure channel [Foschini '85]. The DDA cost function will still have bad local minima. All points with weights such that $\mathbf{T}=\mathbf{G}^T\mathbf{W}$ have an odd number n of elements all equal to $\pm\binom{n-1}{(n-1)/2}2^{-(n-1)}$, the rest being zero. The first several values are 1/2, 3/8 and 5/16 for n = 3, 5 and 7 respectively. A proof is given in Appendix B. [Ding '93] contains a similar proof for PAM data and the Sato algorithm. Using simulation of a seven channel, seven detector system, the existence of the stable points for the n = 3 and 5 case was verified. The stepsize had to be reduced by a factor 3 (from μ=0.06 to μ=0.02) for n=3 and by a factor of 100 (μ=6x10^{-4}) for the n=5 case. This indicates that the region of convergence for the local minima in the DDA cost function is very small. That would explain why we did not notice them during the many simulations done for this research. The stepsize used, 0.06, is, apparently, too large for the region of convergence. Therefore, if **G** is uniquely determined or overdetermined, we have shown that all of the CMA minima completely cancel crosstalk, and for DDA only crosstalk-cancelling minima are provisionally stable, due to step size.

Unfortunately for the crosstalk application, it is not sufficient to guarantee that the output consists of a pure channel, free of crosstalk. Only a specific pair of minima that represent the desired channel and its negative is suitable. All other minima corresponding to other channels represent a failure of the algorithm. In addition, for networks with many channels, we expect that the number of detectors will be smaller than the number of channels. The weights being applied only to a neighborhood of detectors on either side of

the detector receiving the peak of the desired channel. In this case the system will be underdetermined as it is in the equalization problem.

H has a regular structure with each row being a shifted version of the preceding row (it is toeplitz). **G** will only have such a structure if the channel wavelengths and detector locations are regularly spaced and the transfer properties of the demultiplexer are constant with frequency (i.e., the light distribution does not change shape and only shifts at a constant rate across the detector array as a function of frequency). In reality the different laser sources will have different powers and their wavelengths will not be quite regular in spacing, the demultiplexer focusing ability will change with frequency or the detector performance may vary. So, in general, **G** will have a more irregular structure.

The crosstalk data is in parallel vs. the serial nature of the data in the equalizer. In the case of the Bussgang equalizer this should not matter. A Bussgang algorithm will still work if the update rate is divided by N, which would mean that it is working on a completely new batch of data each iteration just as the crosstalk canceller.

4.1.3.DDA and CMA Applied to WDM Crosstalk Cancellation

Standard DDA has a problem with a {0,1} alphabet. The all-zero vector is a global minimum of the DDA cost function adjusted for the OOK data, $J=E\{(y-\frac{1}{2}(\mathrm{sgn}(y-\frac{1}{2})+1)^2\}$. If y is always zero the cost function achieves the global minimum of zero. CMA cannot work with OOK either because the kurtosis of the signal constellation equals two [Agee '88]. Therefore the d.c. value must be removed from each channel by a bank of notch filters for these algorithms to work.

Given a d.c. notch filter, we still must deal with the misconvergence problems of DDA and CMA. In the previous chapter we showed that initial weights and linear constraints chosen without regard to channel gain can still misconverge in DDA and CMA. Weight centering, combined with the CMA for equalizer applications, was shown

to reduce misconvergence by [Focshini '85] and more recently by [Li '95]. We propose a more active weight management technique, called "peak forcing", to be used with both CMA and DDA. Our simulations have shown that for the MMSE weight vector, the element with the maximum magnitude coincides with the detector most aligned with the desired laser wavelength (except for cases of severe crosstalk or laser drift). Motivated by this observation, we suggest an algorithm that monitors the index of the largest-magnitude element and shifts the weight vector to make the largest weight coincide with the proper detector. Zeros are inserted at the end where the shift occurs. For example consider a seven element weight vector with detector four as the desired location for the peak. If, after a weight update, the weights look like $[0.2.\ 0.5,\ 1.2,\ 0.9,\ 0.5,\ 0.3,\ 0.1]^T$, then peak forcing would produce $[0.0,\ 0.2,\ 0.5,\ 1.2,\ 0.9,\ 0.5,\ 0.3]^T$.

In Chapter 5 we describe the results of Monte-Carlo simulations of both DDA and CMA augmented with peak forcing. Both algorithms were successful.

4.2.New Algorithms

4.2.1.Constrained Decision Directed (CDD)

This algorithm is a decision-directed based algorithm designed to work with OOK data (or any modulation scheme where the absence of light is one of the symbols) in a d.c. coupled receiver. It is derived from training-sequence LMS by making two modifications. The first is to update the weights only when a zero is received. If the decision circuitry decides that a zero was received then it can be assumed that S_d is zero and that $E\{e^2\} = E\left\{y^2\right\}$. If S_d is zero, then $\mathbf{G}$ becomes $\mathbf{G}_c$ and $\mathbf{S}$ becomes $\mathbf{S}_c$, where $\mathbf{G}_c$ is $\mathbf{G}$ with the d^{th} column removed and $\mathbf{S}_c$ is $\mathbf{S}$ with the d^{th} element removed (first defined in Chapter 2).

$$E\{e^2\} = E\{y^2\} = E\{(\mathbf{W}^T\mathbf{Z})^2\} = E\{(\mathbf{W}^T(\mathbf{G}_C\mathbf{S}_C+\mathbf{v}))^2\}$$
$$= \mathbf{W}^T(\mathbf{G}_C E\{\mathbf{S}_C\mathbf{S}_c^T\}\mathbf{G}_C^T + E\{\mathbf{v}\mathbf{v}^T\})\mathbf{W} = \mathbf{W}^T(\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_c^T+\sigma^2\mathbf{I})\mathbf{W}, \quad (43)$$

where, again, $\mathbf{R}_{CC} = E\{\mathbf{S}_C\mathbf{S}_c^T\}$, the correlation matrix of the crosstalk vector $\mathbf{S}_C$ and σ^2 is the thermal noise variance. We assume that each element of **S** is an OOK signal with equal probability of zeroes and ones, that all channels have equal power, that the bits of each channel are synchronized and that all of the signals are independent of each other. The equal power assumption is valid since the factors that account for unequal power can be absorbed into the gain matrix **G**. Since the algorithm must operate on zeroes and ones the d.c. component must remain so $\mathbf{R}_{CC}$ is nondiagonal. $\mathbf{R}_{CC}$ has 1/2 on the diagonal and 1/4 elsewhere. An example of a 4 by 4 $\mathbf{R}_{CC}$ is

$$\mathbf{R}_{CC} = \begin{pmatrix} 1/2 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/2 \end{pmatrix}. \quad (44)$$

The second modification is the addition of a linear constraint $\mathbf{L}^T\mathbf{W}=f$, where **L** is the constraint vector and f is a scalar constant. A constraint is necessary for two reasons. The algorithm only updates when $S_d=0$, so $\mathbf{R}_{sd}=\mathbf{0}$. Therefore, without a constraint the optimum weights are **0** (see Equation (8) with $\mathbf{R}_{sd}=\mathbf{0}$). Secondly, the constraint will guarantee that y contains the desired signal S_d.

Again using Lagrange multipliers [Frost '72], the solution is

$$\mathbf{W}_{dd} = (\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_c^T+\sigma^2\mathbf{I})^{-1}\mathbf{L} = k_{dd}(\mathbf{G}_C\mathbf{R}_{CC}\mathbf{G}_c^T+\sigma^2\mathbf{I})^{-1}\mathbf{L}. \quad (45)$$

where k_{dd} is a scalar constant equal to $[\mathbf{L}^T(\mathbf{G}_c\mathbf{R}_{cc}\mathbf{G}_c^T+\sigma^2\mathbf{I})^{-1}\mathbf{L}]^{-1}$.

Comparing Equation (45) to Equation (6) suggests that the best constraint would be $\mathbf{L}=\mathbf{G}_d$, which would make $\mathbf{W}_{dd}$ equal to $\mathbf{W}_{SCNR}$ (except for a scale factor) and SCNR would be optimum. We have assumed that such precise knowledge of the light distribution across the detector array is not available because if it was, the optimum weights could be precomputed and the adaptive algorithm would not be needed. For this reason, we use the simple linear constraint $\mathbf{I}_d\mathbf{W} = 1$, where, again, $\mathbf{I}_d$ is a vector of all zeroes with a one in the d^{th} position. The constraint assumes little about the desired channel light distribution except which detector receives the peak of the desired laser signal.

The SCNR penalty for using $\mathbf{W}_{dd}$ defined in Equation (45) instead of $\mathbf{W}_{MSE}$ (Equation (8)) is less than one dB for many relevant scenarios. Our simulations show that, even though the weights are not absolutely optimum, the SCNR improvements achieved by the CDD algorithm are quite significant.

As described in Chapter 3, a constraint forces the weights that select an incorrect channel to be pushed further away compared to standard Bussgang algorithms that do not use constraints. For example, take a 5 channel WDM network with adjacent channel crosstalk of -12 dB. Table I shows the Euclidean distance between the MMSE weights for channel three and the MMSE weights for the other channels. Note that the distances are greatly increased for the CDD algorithm. Pushing the other MMSE weights away enlarges the region of convergence for channel three and decreases the chances of convergence to other channels. It especially pushes out channels one and five because they have such a great mismatch with the constraint while in the unconstrained DDA case channels 1 and 5 are actually the closest.

Table I. Euclidean distance of MMSE weights from MMSE weights of channel 3

	Channel 1	Channel 2	Channel 4	Channel 5
Decision Directed	1.3	1.8	1.8	1.3
Constrained Decision Directed	15.3	3.6	3.6	15.3

CDD will adaptively converge to $\mathbf{W}_{dd}$. The algorithm is simply:

Form new output: $y = \mathbf{W}_{old}^{T}\mathbf{Z}$

Decide if y is a one or a zero

If a one then coast $\mathbf{W}_{new} = \mathbf{W}_{old}$

If a zero then update $\mathbf{W}_{new} = \mathbf{W}_{old} - \mu y\mathbf{Z}$

and then set $\mathbf{W}_{new} = \mathbf{W}_{new} / \mathbf{W}_{new}^{(d)}$

Repeat

$\mathbf{W}_{new}^{(d)}$ is the weight of the d^{th} detector. The last step implements the linear constraint $\mathbf{I}_d^T\mathbf{W} = 1$. Figure 18 shows a block diagram of the CDD algorithm using the architecture of the generalized sidelobe canceller (see Chapter 2).

The CDD algorithm has some drawbacks. It requires that the output of each detector be sampled within a single bit interval because the algorithm only runs when a zero is received. Since the bit rate may be in the GHz range, the sampling electronics could be expensive to implement. The nature of the algorithm will not permit the algorithm to find the desired channel if the laser drifts too far from the assumptions built into the linear constraint. Finally the algorithm will work only with a desired channel transmitting an OOK signal (or at least an alphabet with one symbol represented by the absence of light). Its main advantage is that it is a completely blind adaptive algorithm that

determines the proper weights solely by examining the output of the channel, relieving the network of the overhead due to training sequences.

4.2.1.1.Relationship of CDD Algorithm to VC_1

CDD is closely related to VC_1, described in the previous chapter. Specifically, if VC_1 is modified to operate with OOK data, its stable points are identical to CDD. For proof, we will examine the cost function of VC_1 modified for use with OOK: For standard VC_1 the algorithms must decide whether to feedback $+E\{|y|\}$ or $-E\{|y|\}$. For a $\{0,1\}$ alphabet the decision is between zero or $2E\{y\}$. The latter value takes the place of "1" and allows the channel gain estimate to be incorporated into the algorithm. The optimum decision threshold is $E\{y\}$ [Proakis '89] so the "decision" function must switch from zero to $2E\{y\}$ at a threshold of $E\{y\}$, $f_d = \left(\text{sgn}\left(y - E\{y\}\right) + 1\right)E\{y\}$ will suffice giving a cost function of,

$$J_{VC}^{1} = E\left\{\left[y - f_d(y)\right]^2\right\} = E\left\{\left[y - \left(\text{sgn}\left(y - E\{y\}\right) + 1\right)E\{y\}\right]^2\right\}. \tag{46}$$

To convert to an algorithm that only runs on zeros we must add an additional "switch" to the cost function to eliminate all the cases that exceed the threshold. $f_s(y) = \frac{1}{2}\left(1 - \text{sgn}\left(y - E\{y\}\right)\right)$ equals 1 at zero and switches to 0 at $E\{y\}$. Adding this second switching function gives the CDD cost function:

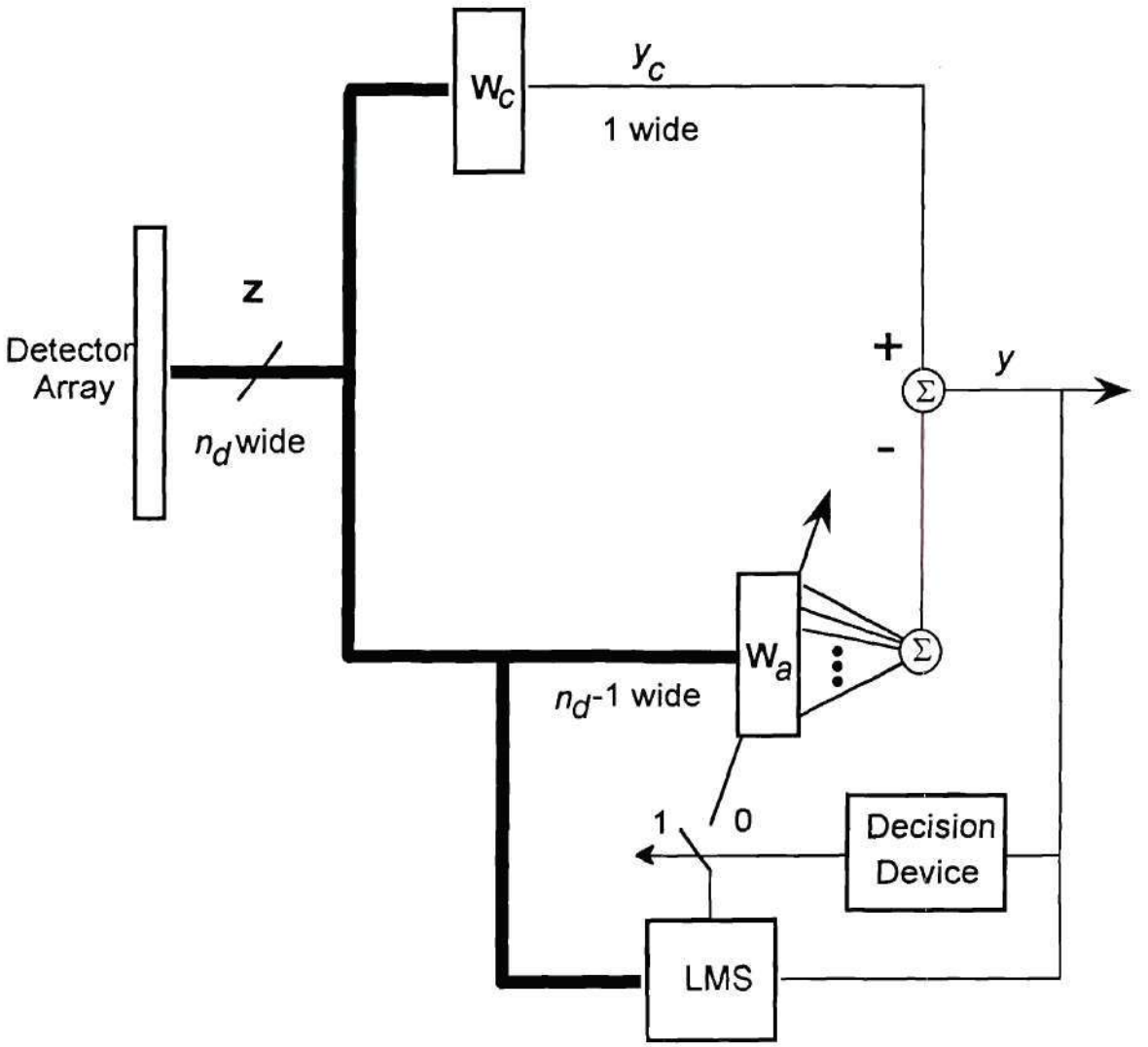


Figure 18: Block diagram of the CDD algorithm.

$$
\begin{aligned}
J_{CDD} &= E\left\{\left[y-\left(\operatorname{sgn}\left(y-E\{y\}\right)+1\right)E\{y\}\right]^2\frac{1}{2}\left(1-\operatorname{sgn}\left(y-E\{y\}\right)\right)\right\} \\
&= \frac{1}{2}E\left\{\left[y-\left(\operatorname{sgn}\left(y-E\{y\}\right)+1\right)E\{y\}\right]^2\right\}-\frac{1}{2}E\left\{\left[y-\left(\operatorname{sgn}\left(y-E\{y\}\right)+1\right)E\{y\}\right]^2\operatorname{sgn}\left(y-E\{y\}\right)\right\}, \quad (47) \\
&= \frac{1}{2}J_{VC}^{1}+0
\end{aligned}
$$

where the second term evaluates to zero because of the symmetry of the distribution of y about its mean. So we see that updating only on zeros leaves the cost function unchanged except for a scale factor.

The cost function can be manipulated in another way. Return to Equation (47) and observe that only the realizations of the quantity within the expectation that are less than the threshold, $E\{y\}$, have nonzero values. In this case, the value in the expectation reduces to y^2 so another way to rewrite the cost function is as a conditional expectation of the random quantity y^2,

$$
J_{CDD} = \frac{1}{2}E\left\{y^2 \mid y < E\{y\}\right\}. \quad (48)
$$

The gradient of Equation (48) is $\nabla_{\mathbf{W}} J_{CDD} = E\left\{\mathbf{Z}y \mid y < E\{y\}\right\}$, which directly implies the CDD algorithm defined above. Figure 19 shows a contour plot of the cost function using the example channel and equalizer from Chapter 3. This plot will apply to both CDD and VC_1 modified for OOK. Note that the contours no longer have symmetry about the origin. This is due to the non symmetrical symbol constellation.

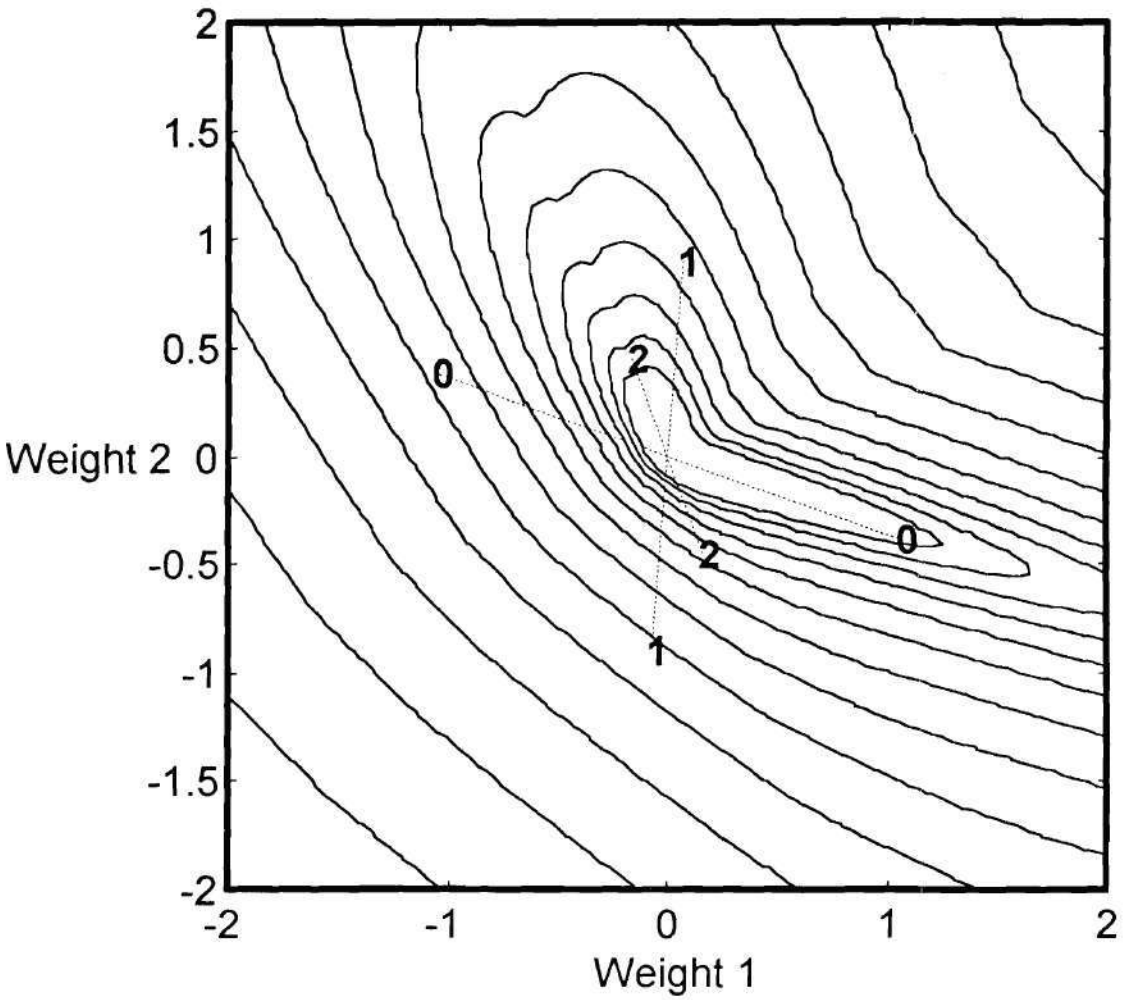


Figure 19: Contour plot of CDD cost function (h_m={1.0, 0.42}).

4.2.2.Pilot Tone Algorithm

Several authors [Way '92, Poggiolini '94] have proposed that each transmitter in a WDM network add a sine wave of unique frequency to the laser drive current for each channel. The sine wave frequencies are higher than the bandwidth of the information so they can be isolated and used for channel identification, signal routing and other uses. We propose to use these pilot tones to adaptively cancel the linear crosstalk and compensate for laser frequency drift. We propose two methods, an adaptive unconstrained LMS algorithm and a method that uses simple filters to estimate the crosstalk gain matrix **G** ([Ho '95] independently proposes the latter technique).

The pilot tone algorithm works as follows: each transmitting laser in the network adds a small sinusoidal oscillation to the laser drive current. The drive current is usually the sum of a bias current and a current containing the information; now we add a third current term that is a sine wave of constant frequency. The frequency must be higher than any spectral component of the information current. The resulting spectrum would look like Figure 20. Note that the pilot tone is separated from the data spectrum and can be isolated with a band pass filter.

Figure 21 shows the same spectrum with data contaminated with crosstalk from other channels, each with their own pilot tones of unique frequency. We assume that the frequencies of the pilot tones are small compared to the differences between the various channel wavelengths. If so, the pilot tones will be subject to essentially the same crosstalk gains (**G**) as the information. Therefore, if weights are found to cancel the crosstalk among the pilot tones, the same weights will cancel the crosstalk in the data.

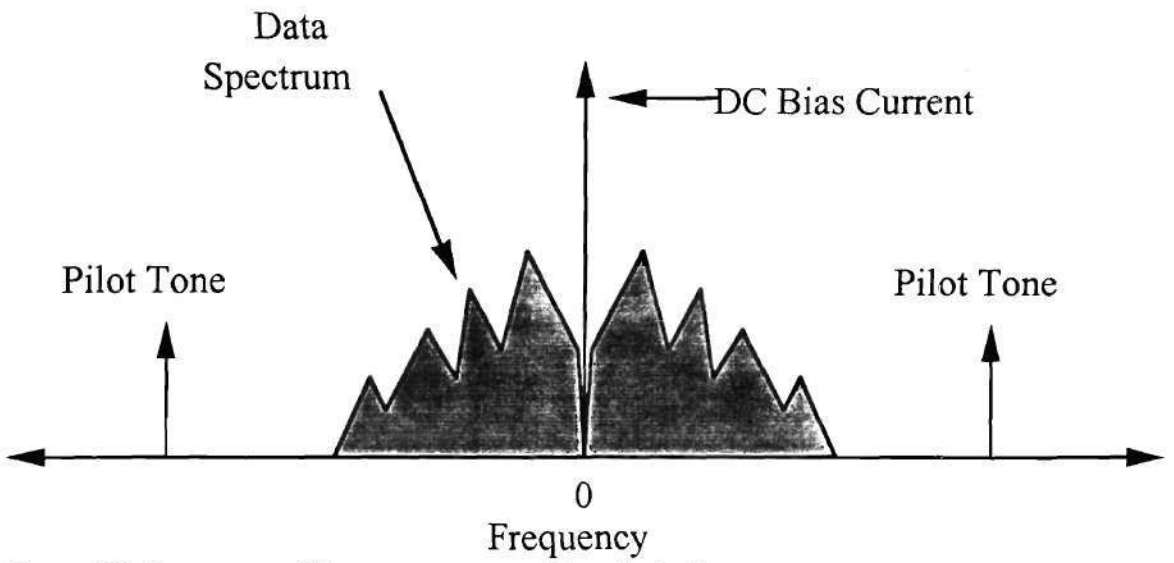


Figure 20: Spectrum of detector current with added pilot tone.

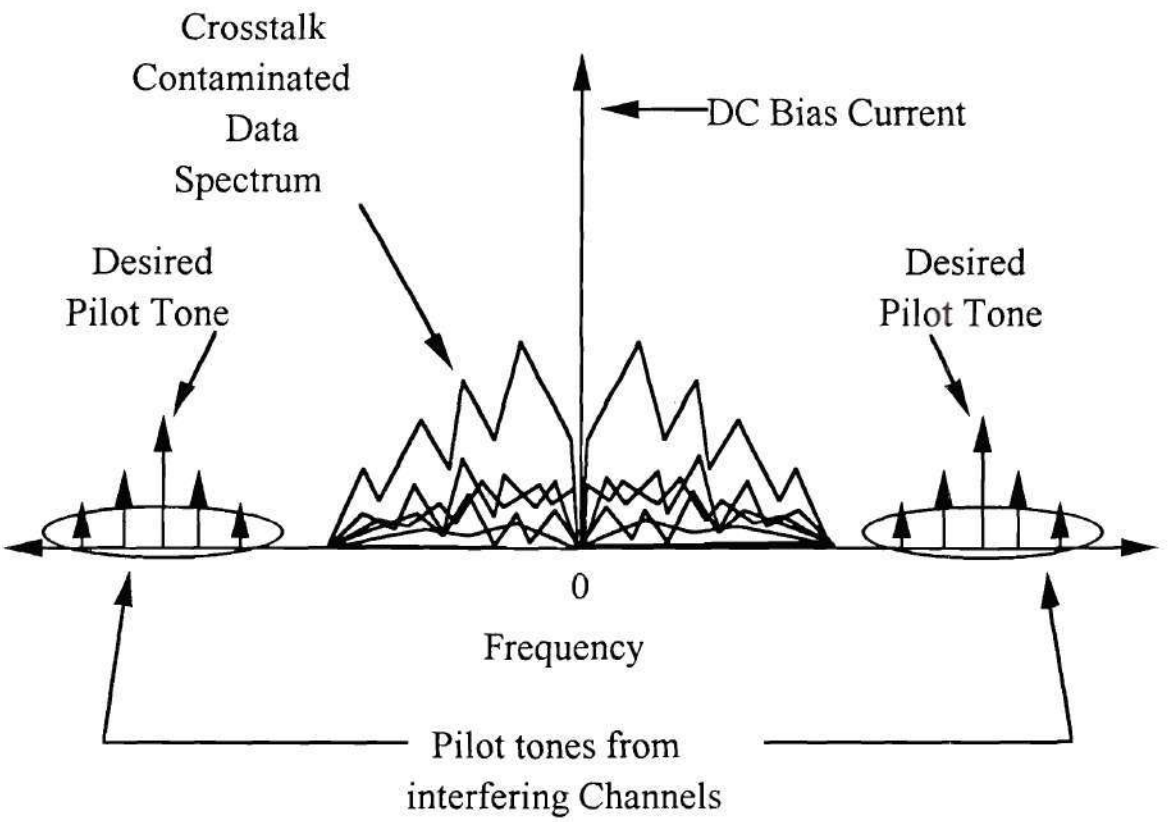


Figure 21: Spectrum of dectector current of desired laser signal with crosstalk from other channels.

Figure 22 shows a schematic diagram of the pilot tone LMS receiver. A bandpass filter isolates the entire field of pilot tones and a lowpass filter isolates the data signals. The pilot tones can then be beat down with an optional local oscillator. An unconstrained LMS algorithm is then run on the field of pilot tones. Unconstrained LMS needs a desired signal, which in this case is a sine wave at the frequency of the pilot tone of the desired channel. The resulting weights that are generated are then applied to the data signals to eliminate the crosstalk.

Using pilot tones to cancel crosstalk offers several advantages.

- *No need for high speed electronics to perform the adaptive algorithm*: Theoretically, the sine waves may be arbitrarily close in frequency. So the adaptive filter, whether implemented digitally or analog, can run as slowly as desired. In other words, the network designer is free to trade off convergence speed and cost.

- *Unconstrained LMS algorithm*: Because the signals are now simple sine waves, it is easy to synthesize a desired signal; it need only be a sinusoid of fixed amplitude at the proper frequency and close to the correct phase. This allows the simpler unconstrained LMS algorithm to be used. The desired signal can be generated using a variety of methods, most of them using the input signal. Rough knowledge of the desired pilot tone frequency is needed then the pilot tone can be electronically filtered out and used to lock the desired signal to the proper phase and frequency.

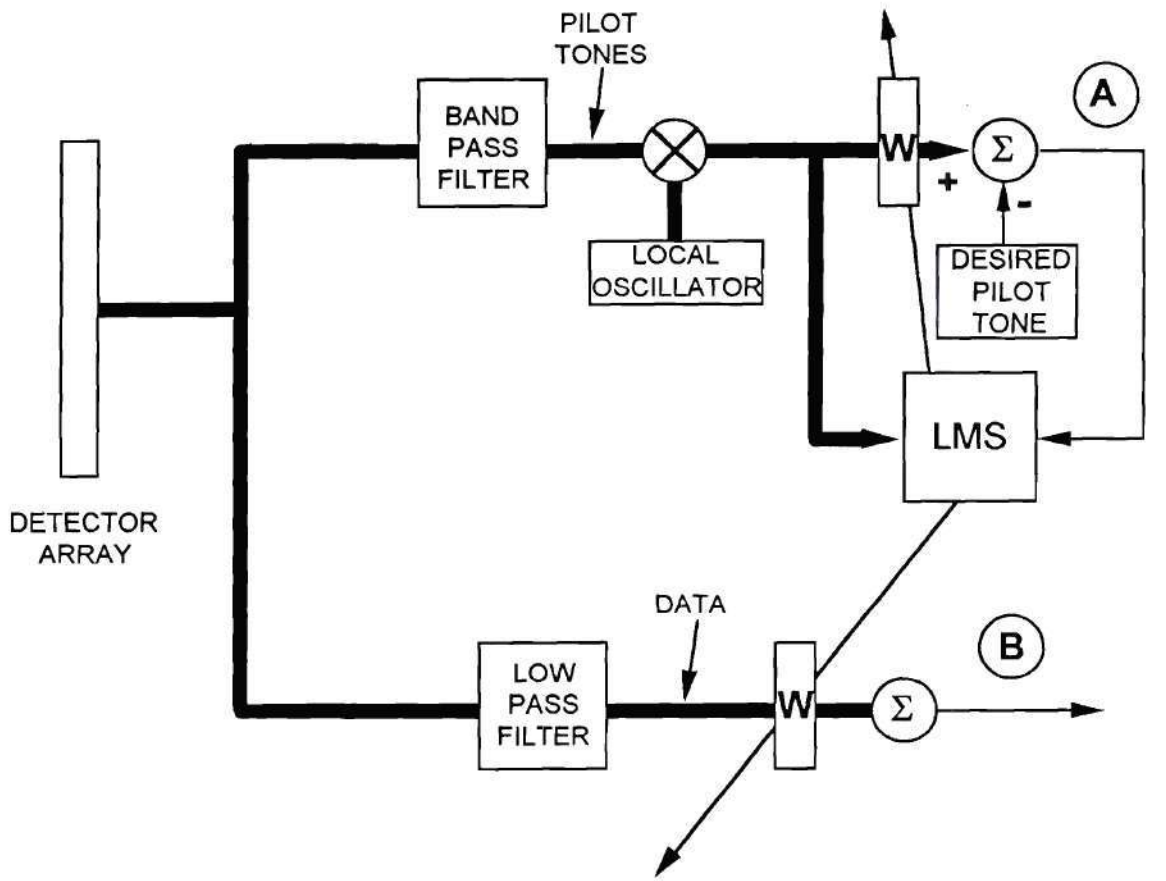


Figure 22: Block diagram of pilot tone receiver.

- *Tracking of laser drift and/or changes in the demultiplexer*: An added benefit of the unconstrained LMS is that the weights are automatically configured to follow a drifting laser or to account for changes in the demultiplexer performance (e.g., due to temperature effects). This is because the pilot tones are part of the intensity modulation and therefore are independent of the laser center frequency. In fact the receiver need absolutely no knowledge of the actual laser frequencies; all it needs to know are the pilot tone frequencies assigned to each channel. The CDD does not have this property because some assumptions on the location of the desired-laser energy on the detector array are built into the constraint. Note that if a drifting laser coincides with another laser center frequency no linear combining technique can separate the channels.

- *Data Format independence*: Pilot tone cancellation will improve the signal to crosstalk ratio no matter what types of data are being transmitted on the lasers. It will work with any mixture of analog or digital intensity-modulated data.

4.2.3.Optimum Weights using Pilot Tone LMS

We will begin by comparing the optimum weights for both algorithms to see how well the optimum pilot tone weights cancel crosstalk when they are applied to the actual data. The equation for the optimal pilot tone weights is the same as the optimal MSE weights in Equation (8) with the signal vector **S** replaced by the pilot tone vector **P**

$$\mathbf{W}_p = (\mathbf{G}E\{\mathbf{P}\mathbf{P}^T\}\mathbf{G}^T + \sigma^2\mathbf{I})^{-1}\mathbf{G}\,E\{d\mathbf{P}\} = (\mathbf{G}\mathbf{R}_{pp}\mathbf{G}^T + \sigma^2\mathbf{I})^{-1}\mathbf{G}\mathbf{R}_{pd}. \qquad (49)$$

$\mathbf{R}_{pp}$ is a diagonal matrix and $\mathbf{R}_{pd}$ is a multiple of the d^{th} column of $\mathbf{R}_{pp}$, a vector with one non-zero entry in the d^{th} position. $\mathbf{R}_{ss}$ is also diagonal (assuming the d.c. bias

current has been filtered out) so if we assume that the ratio between the power of the pilot tone and the data signal is the same in each channel then $\mathbf{R}_{pp}$ and $\mathbf{R}_{pd}$ differ from and $\mathbf{R}_{ss}$ and $\mathbf{R}_{sd}$ by only a scale factor. Therefore, the only difference between the weights $\mathbf{W}_p$ and $\mathbf{W}_{MSE}$ would be due to differences in the SNR between the two cases. As long as the SNR is high the differences should be negligible.

$\mathbf{R}_{pp}$ and $\mathbf{R}_{ss}$ may differ by more than a scale factor. The pilot tone to data power ratios may not be constant for all channels or the filters that separate the data and pilot tones may not be matched from channel to channel and could alter the power ratios. Since we are forming the weights using $\mathbf{R}_{pp}$ and applying them to the data we are interested in the question of how the weights vary as a function $\mathbf{R}_{ss}$. For this analysis we consider three types of receivers: 1) the same number of detectors as channels (uniquely determined), 2) more detectors than channels (overdetermined) and 3) more channels than detectors (underdetermined). For the first two cases we show that when receiver noise is neglected the pilot tone and training-sequence LMS weights for the d^{th} channel are equal to the d^{th} column of the matrix $(\mathbf{G}^T)^{\#}$ where $^{\#}$ denotes the Penrose pseudo-inverse (if the matrix is invertible the pseudo-inverse is equal to the inverse). When receiver noise cannot be neglected, the excess MSE produced by using pilot tone weights instead of training-sequence weights is proportional to the square of the noise power (σ^4) while the residual MSE produced by using optimum weights is proportional to the noise power (σ^2).

Uniquely determined: The simplest case is a receiver with the same number of channels and detectors, for example seven channels and seven detectors. In this case $\mathbf{G}$ is square. If we neglect thermal noise, then $\mathbf{R}_{ZZ} = \mathbf{G}\mathbf{R}_{ss}\mathbf{G}^T$ and $\mathbf{R}_{ZZ}$ has full rank and is invertible. The solution to the Wiener-Hopf equation is then

$$\mathbf{W}_{MSE} = (\mathbf{G}\mathbf{R}_{ss}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{R}_{sd}. \tag{50}$$

Recall that $\mathbf{R}_{sd}$ is simply a multiple of the d^{th} column of $\mathbf{R}_{ss}$. We may solve for the optimum weights for all channels simultaneously; then $\mathbf{R}_{sd}$ is a square matrix equal to $\mathbf{R}_{ss}\mathbf{A}$ where $\mathbf{A}$ is a diagonal matrix representing the fact that each column of $\mathbf{R}_{sd}$ is a scalar multiple of the corresponding column of $\mathbf{R}_{ss}$. The solution to the Wiener-Hopf equation can now be expressed as

$$\mathbf{W}_{MSE} = (\mathbf{G}\mathbf{R}_{ss}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{R}_{ss}\mathbf{G}^T(\mathbf{G}^T)^{-1}\mathbf{A} = (\mathbf{G}^T)^{-1}\mathbf{A}, \tag{51}$$

where the term $\mathbf{G}^T(\mathbf{G}^T)^{-1}$ can be inserted because it is the identity matrix. Therefore each column of $(\mathbf{G}^T)^{-1}$ is the optimal weight vector for selecting the corresponding channel. The important thing to observe is that the optimal weights are independent of $\mathbf{R}_{ss}$ as long as it is nonsingular, which is the case for practical situations where each data channel is uncorrelated. Therefore weights derived from pilot tones will be optimal for all types of WDM networks whether analog, digital or hybrid networks containing any combination of data types. In the rest of the thesis the diagonal matrix will be dropped because it only changes the weights by a scale factor and doesn't affect SIR.

The case just described is the simplest because all matrices are square and invertible. Two other cases are also considered: the overdetermined case, where the number of detectors exceeds the number of WDM channels, and the underdetermined case where the number of detectors is less than the number of channels. The overdetermined case may occur if laser drift is perceived to be a problem. The detectors may be placed more closely in frequency than the WDM channels to make sure that the laser light is received even if the channels drift from their assigned wavelengths. For example, the detectors may be spaced at one nm intervals while the channels have a nominal two nm spacing. For this case, the crosstalk matrix $\mathbf{G}$ would be "tall" having more rows than

columns. The underdetermined case might occur if the WDM network has many channels. It is assumed that the crosstalk problem is caused by the channels nearest in frequency and that the crosstalk from channels far away in frequency can be neglected. In this case, the weights are applied to only the several adjacent channels and the matrix **G** will be "flat," having more columns than rows. The derivations for the following results are given in the Appendix **C**.

Overdetermined: For this case, we get

$$\mathbf{W}_{\mathrm{MSE}} = (\mathbf{G}^{T})^{\#}. \tag{52}$$

This is the same result as in the invertible case except that the pseudo-inverse replaces the actual inverse. Again note that the optimum weights do not depend on $\mathbf{R}_{\mathrm{SS}}$ so the pilot tone weights would be optimum for the data also.

Underdetermined: For this case, $\mathbf{W}_{\mathrm{MSE}}$ cannot be made independent of $\mathbf{R}_{\mathrm{SS}}$. All that can be said is

$$\mathbf{W}_{\mathrm{MSE}}\mathbf{P}_{\mathrm{G}} = (\mathbf{G}^{T})^{\#}, \tag{53}$$

where $\mathbf{P}_{\mathrm{G}}$ is the projection operator onto the space spanned by the rows of **G**. So the optimum weights do vary with $\mathbf{R}_{\mathrm{SS}}$, however the projection of the weights into the space spanned by the rows of **G** is independent of $\mathbf{R}_{\mathrm{SS}}$ and is equal to $(\mathbf{G}^{T})^{\#}$. Therefore, if $\mathbf{R}_{\mathrm{pp}} \neq \mathbf{R}_{\mathrm{SS}}$, $\mathbf{W}_{\mathrm{p}}$ for the underdetermined case may be suboptimum for the data. However, if the extra channels are so far away in frequency that their crosstalk can be considered negligible then their effects on the optimum weight vector should be very small.

Calculation of optimum weights for simulated networks using randomly generated $\mathbf{R}_{ss}$ support this idea.

In summary, for the uniquely determined and overdetermined cases, $(\mathbf{G}^T)^{\#}$ yields the lowest MSE, regardless of the data statistics, $\mathbf{R}_{ss}$. For the underdetermined receiver, the pilot tone weights generate weights that are nearly optimal.

When noise is reintroduced the optimum weights become a function of $\mathbf{R}_{ss}$. Noting that $\mathbf{R}_{zz}=(\mathbf{G}\mathbf{R}_{ss}\mathbf{G}^T+\sigma^2\mathbf{I})$, the optimum weights are

$$\mathbf{W}_{MSE} = (\mathbf{G}^T)^{-1} - \sigma^2\mathbf{R}_{zz}^{-1}(\mathbf{G}^T)^{-1}. \tag{54}$$

For the overdetermined case, we get

$$\mathbf{W}_{MSE} = (\mathbf{G}^T)^{\#} - \sigma^2\mathbf{R}_{zz}^{-1}(\mathbf{G}^T)^{\#}. \tag{55}$$

For the underdetermined case, we can only say

$$\mathbf{W}_{MSE}\mathbf{P}_G = (\mathbf{G}^T)^{\#} - \sigma^2\mathbf{R}_{zz}^{-1}(\mathbf{G}^T)^{\#}. \tag{56}$$

For each case, the matrix of optimum weights, or its projection, consists of an invariant term that does not depend on $\mathbf{R}_{ss}$ plus a term that does depend on $\mathbf{R}_{ss}$ that is scaled by the noise power.

The final step is to examine ΔJ, the excess MSE produced by using $\mathbf{W}_p$ instead $\mathbf{W}_{MSE}$. If we expand ΔJ into a power series and take only the first term we get

$$\Delta J \approx \sigma_t^4 \mathrm{Tr}[(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{R}_{ss}^{-1}(\mathbf{I}-(\sigma_p/\sigma_t)^2\mathbf{R}_{ss}\mathbf{R}_{pp}^{-1})^2(\mathbf{G}\mathbf{G}^T)^{-1}], \tag{57}$$

where the subscripts t and p signify training-sequence and pilot tones respectively and Tr(•) is the trace operator (the sum of the elements of the main diagonal of the matrix). Note that the excess MSE is proportional to the square of the noise power. Contrast this to J_{min}, which is the MSE due to the optimum weights. If the expansion is taken to two terms (the same accuracy used above) then the minimum error, using optimum $\mathbf{W}_{MSE}=\mathbf{R}_{zzt}^{-1}\mathbf{R}_{zd}$, is

$$J_{min} \approx \sigma_t^2 \mathrm{Tr}((\mathbf{G}\mathbf{G}^T)^{-1}). \tag{58}$$

We see that J_{min} is proportional to the noise power, so for reasonably high SNR, say greater than 20 dB, the excess MSE will be negligible compared to the minimum MSE. The above analysis is not valid for an underdetermined receiver (**G** is "flat") however the weights should not be very different if the uncompensated channels contribute very little crosstalk.

4.2.4.Direct Estimation of the gain matrix G

The analysis in the preceding section shows that $(\mathbf{G}^T)^{\#}$ makes an excellent approximation of the optimum weights in all cases.[2] This suggests that if **G** could be directly measured or estimated then the optimum weights could be computed directly. The pilot tone concept allows direct estimation of **G**, the elements of which can be measured with simple filters. Consider the shape of the spectrum of pilot tones in Figure 21. The strength of each spectral line is proportional to the amount of light coupled into that

---

[2] The only time they would fail to do so would be if $\mathbf{R}_{ss}$ does not have full rank, which should never occur. Even in this unlikely case $(\mathbf{G}^T)^{\#}$ would still do an excellent job of eliminating the crosstalk; it would just mean that another weight vector could do a slightly better job.

detector from a channel. In other words the strength of the various pilot tones compose a row of the gain matrix **G**. Passing the pilot tone spectra from each detector through a filter bank and measuring the outputs will determine the entire matrix **G**. The filter outputs could be monitored and the weights recomputed whenever **G** changes. The measurements could be done with either analog or digital filters.

Instead of filters a fast Fourier transform (FFT) could be performed and some sort of threshold detection be used to find the elements of **G**. This method may have an advantage because the number of weights used for cancellation could be adjusted as network conditions change. For example, find the detector with the maximum desired pilot tone then find all other pilot tones in that detector that exceed a threshold. Next, find which detectors in the array have the maximum values for each interfering pilot tone. Only these detectors would be used for crosstalk cancellation. In this way a minimum set of detectors would be used.

4.2.5.Data format independence

The pilot tone algorithm is independent of data format, i.e., it will work with any mix of digital and analog formats. Figure 23 shows the result of a simulation of the pilot tone LMS with a seven channel network that illustrates this property. Channel 4 carries OOK digital data and channel 5 a simple analog signal. Figure 23a and Figure 23b shows the output of the filter (marked "B" in Figure 22) tuned to channel 4 and 5 respectively. Detector and channel spacings were 2 HWHM. Note that the filter converges for both types of data.

desired channel, nothing can be done because the linear crosstalk power from the drifting laser will be too large to cancel. However, the drifting channel may collide with another undesired channel. In that case, the two channels may beat together to create significant nonlinear crosstalk energy.

The worst case would be when the laser for a channel twice removed from the desired drifts towards the desired channel and collides with an adjacent channel. For example, assume that channel 4 is desired and that channel 2 drifts in frequency and collides with channel 3 creating a nonlinear beating signal. This scenario would couple the largest amount of beating signal into the detector for the desired channel.

4.3.1.2.Derivation of nonlinear beating terms

Assume that a "rogue" laser drifts in frequency until it nearly coincides in frequency with an adjacent laser so that a beating term is passed by the electronics. In order to analyze how this beating term impacts the algorithms we need to determine what the beating term looks like as a function of time and how the beating term is distributed across the detector array. The beating term, S_b, is added to the signal vector **S** and a vector representing the distribution of the beating term $\mathbf{G}_b$, becomes a corresponding additional column of the crosstalk matrix **G**. In other words, treat the beating term as if it is an additional channel to be cancelled.

The light amplitude for the n^{th} laser channel is

$$a_n = S_n(t)\cos(\omega_n t + \varphi_n(t)), \tag{59}$$

where $S_n(t)$ is the modulating signal of the n^{th} channel, ω_n is the laser frequency, and $\phi_n(t)$ is a laser phase noise term that is uncorrelated from channel to channel. The rogue channel amplitude (a_r) and the "victim" channel (a_v) are now very close in

frequency so $\omega_r = \omega_v + \Delta\omega$, and $\Delta\omega$ is small enough to be passed by the receiver electronics. The beating term arises when a_r and a_v are multiplied together as follows:

$$\begin{aligned} S_b &= S_r(t)S_v(t)\cos(\omega_v t + \Delta\omega + \varphi_r(t))\cos(\omega_v t + \varphi_v(t)) \\ &= \tfrac{1}{2}S_r(t)S_v(t)\big(\cos(2\omega_v t + \Delta\omega + \varphi_r(t) + \varphi_v(t)) + \cos(\Delta\omega + \varphi_v(t) - \varphi_v(t))\big). \end{aligned} \quad (60)$$

The term with the sum of the frequencies will be eliminated by the electronics since ω_v is on the order of 10^{15} rad/s, so

$$S_b = \tfrac{1}{2}S_r(t)S_v(t)\cos(\Delta\omega t + \Delta\varphi(t)) \quad (61)$$

is appended to the end of vector **S**.

The elements of the crosstalk gain matrix **G** describe how the channel intensities are distributed across the detector array. Therefore the amplitude distribution can be described with terms equal to the square root of the elements of **G**. The beating signal S_b is distributed across the detector array by

$$g_{ib} = \sqrt{g_{ir}}\sqrt{g_{iv}}\,. \quad (62)$$

In other words, each element of $\mathbf{G}_b$ is the geometric mean of the corresponding elements of $\mathbf{G}_r$ and $\mathbf{G}_v$. The vectors $\mathbf{G}_r$ and $\mathbf{G}_v$ will be nearly identical because their laser wavelengths are nearly identical, differing by only a few hundredths of a nm. Consequently, $\mathbf{G}_b$ will also be nearly the same. For example, given a demultiplexer with a one nm HWHM and an array of five detectors spaced at intervals of two nm, and λ_r and λ_v separated by 0.04 nm ($\Delta\omega$=5.3 GHz) produced the following detector distributions:

$$\mathbf{G}_r = \begin{pmatrix} 0.0044 \\ 0.2641 \\ 0.9995 \\ 0.2364 \\ 0.0035 \end{pmatrix} \qquad \mathbf{G}_v = \begin{pmatrix} 0.0039 \\ 0.2500 \\ 1.0000 \\ 0.2500 \\ 0.0039 \end{pmatrix} \qquad \mathbf{G}_b = \begin{pmatrix} 0.0041 \\ 0.2570 \\ 0.9997 \\ 0.2431 \\ 0.0037 \end{pmatrix}. \tag{63}$$

Notice that they are nearly identical so that a single degree of freedom in weight space can do a good job of canceling all three signals. This is fortunate in the case of the pilot tone algorithm because the nonlinear signal may not be present in the pilot tone region of the spectrum so that the LMS algorithm cannot directly cancel it. The results of a simulation of this scenario is described in the following chapter.

4.3.2.Laser drift

Over time the laser frequencies may drift due to temperature effects and component aging. In addition, the frequency response of a grating-based demodulator may vary with temperature. The response of the CDD and pilot tone algorithms to laser drift has been simulated. The pilot tone algorithm converges to the optimum weights in all cases, while the CDD algorithm only succeeds while the desired laser remains between the two neighboring channels i. e., as long as the desired channel has the best match to the constraint. Details of the simulation are presented in the next chapter.

The fully blind algorithms (i.e., all but the PT algorithms) cannot handle unlimited drift; however, strategies may be developed that will allow the receiver to measure small amounts of drift after they have successfully locked on. This knowledge would allow the algorithm to alter constraints, or other parameters that would help the algorithm to lock on the next time it is initialized. The use of such meta-strategies may allow the algorithms to follow large amounts of laser drift that may accumulate over a long time.

For the CDD algorithm the constraint vector can be updated to be approximately equal to the optimum constraint $\mathbf{H}_d$ by computing $\hat{\mathbf{H}}_d = \mathbf{R}_{zz}\mathbf{W}$. The approximation is good if $\mathbf{W}$ produces high SINR (see Equation (42)). The next time the receiver tries to acquire channel d use $\hat{\mathbf{H}}_d / \| \hat{\mathbf{H}}_d \|$ as the constraint (the normalization keeps the magnitude of the constraint equal to one). If the channel drifts over a long period of time the constraint will be updated to assure successful acquisition. Note that $\mathbf{R}_{zz}$ must be estimated with the d.c. component of the signals removed. At this time we have no meta-strategies for DDA and CMA augmented with peak forcing.

CHAPTER 5.

SIMULATION OF CROSSTALK CANCELLATION

All crosstalk cancellation algorithms were tested with a computer simulation. First, a simple model of a grating demultiplexer was created. The demultiplexer has a Gaussian "impulse response", which means that monochromatic light emerging from the fiber forms a spot with a Gaussian intensity profile across the detector array. Changing laser frequency causes the spot to move across the detector array, however the shape of the profile remains constant. The intensity profile does not decay below -50 dB in order to simulate unintentional diffraction from the grating. The profile is normalized so that the peak is one and the half-width-half-maximum (HWHM) is equal to one. HWHM is the distance between the peak of the profile and the -3 dB point on the skirt of the electrical intensity profile, which should be roughly equivalent to the resolution of a grating demultiplexer. Channel spacing and detector array spacing in the simulation is done in units of HWHM. This way any grating demultiplexer can be simulated if one has knowledge of its measured or specified performance, e.g., a given crosstalk level of -12 dB between channels is simulated by specifying detector and channel spacings of two HWHM because this places a detector at the 0.25 point on the Gaussian curve. The formula for the normalized optical intensity profile is

$$i(x) = \exp(-(0.5887x)^2) + 10^{-5}. \tag{64}$$

The laser bandwidth and modulation rate are considered to be small relative to the demultiplexer passband so the channel spectra is not broadened relative to the impulse profile.

In all but one simulation only linear crosstalk is considered; this is reasonable if the channel spacing is much greater than the electronic bandwidths of each IM/DD receiver and the optical powers are not high enough to cause nonlinear effects like stimulated Raman scattering (see previous chapter). One simulation included square law nonlinearities in order to test its effects on the algorithms. Digital signals are assumed to be OOK and are used in all simulations except for a pilot tone simulation where some of the channels were intensity-modulated analog signals (to test the ability of the algorithm to operate in mixed format networks). Inter-symbol interference is not modeled. All signals, including pilot tones, have unit amplitude. The dominant noise source is assumed to be receiver noise that is modeled as a vector of additive zero mean white Gaussian noise. The noise power is equal in all detectors.

The bandpass and lowpass filters shown in Figure 22 were assumed to be ideal so the pilot tone signals and the data signals are perfectly separated. Finally, the stochastic algorithm used an adaptive multidimensional step size and a discrete cosine transformation of the input data to speed convergence [Goldstein '93, Minardi '93].

DDA and CMA were augmented with the peak forcing algorithm and tested against 2000 randomly generated scenarios. The scenario had seven channels and seven detectors with channel four the desired. Both detectors and channels had a nominal separation of two HWHM. The detector location and channel wavelengths were randomly varied with a uniform distribution of ±½ HWHM in order to simulate laser drift or detector misalignment. The d.c. notch filter was assumed to be ideal so the signal vector **S** was made up of {-1,1} signals. The initial weights were totally random except that $||\mathbf{W}_0||$ was set equal to one. For all cases, the DDA algorithm converged to the desired channel four. For the CMA algorithm 37 cases settled into a limit cycle. It seemed that the boundary defined by the peak forcing algorithm was within the region of convergence of an

undesired channel. As the weights moved toward the optimum weights for the undesired channel they crossed the boundary and the peak forcing algorithm would adjust them back across the boundary. This scenario repeated itself forming a limit cycle. We adjusted the peak forcing algorithm so that it would only operate once every 300 iterations of the CMA algorithm. The hope was to give the weights time to get close to the optimum weights for the undesired channel. Then, when the peak forcing does occur, the weights are thrown into the region of convergence of the desired channel, thus, preventing a limit cycle. This proved successful and all of the remaining scenarios converged to channel four using the modified peak forcing algorithms.

Simulated CDD filter output is shown in Figure 24 for a case with 9 cancellation weights and 1.6 HWHM channel and detector separation. The network has 19 channels with channel 10 the desired. The algorithm was initiated at iteration 200. Receiver noise power was set at -50 dB to ensure that the network was crosstalk-limited rather than noise limited. Figure 24a is the filter output and Figure 24b is the output SCNR. With no cancellation, SCNR is 1.5 dB. After convergence it is about 31.5 dB. The SCNR using optimum weights from Equation (45) is 32 dB.

Figure 25 shows the required channel spacing as a function of SCNR. Again, a 19 channel system is modeled with the center channel, 10, as the desired. The curves represent the SCNR performance achieved using optimal MSE weights. The curves are plotted for no cancellation, 3 weights, 5 weights, and 9 weights. Note the large improvement that can be gained by just using two additional detectors to cancel the adjacent channels. Figure 26 recasts the same data by showing how many channels could be fit in the 200 nm wide 1550 nm low loss fiber window for a given SCNR requirement. For this figure, a 1 nm HWHM is assumed for the demultiplexer. The symbols in Figure 25 and Figure 26 are simulation results for both the pilot tone and CDD algorithms. Note that

both the pilot tone and CDD converge to within a fraction of a dB of the optimum for all but very close channel spacings. DDA and CMA are not shown on this figure however they both show the same near-optimal performance.

These figures demonstrate the potential of postdetection crosstalk cancellation to greatly increase the capacity of WDM networks. The required channel spacing to achieve a desired SCNR is reduced by about 30%, 45%, and 60% for 3, 5 and 9 weights respectively.

Over time the laser frequencies may drift due to temperature effects and component aging. In addition the frequency response of a grating-based demodulator may vary with temperature. Because the pilot tone LMS algorithm runs unconstrained, no assumptions are needed on which detectors are, or are not receiving the desired signal energy. Generation of a sine wave at the proper frequency will cause the weights to configure to pick up the desired channel no matter where it is. CDD can accommodate limited laser drift. However, if the desired laser moves so far that another channel is a better match to the constraint vector, the algorithm will lock on to it instead of the desired signal. Figure 27 shows the response of each algorithm to a drifting laser; again, a seven channel, two HWHM network was simulated. The nominal placement of channels one through seven is [-6, -4, -2, 0, 2, 4, 6]. We moved the center frequency of the desired channel (channel 4) to several points zero and three HWHM. At each point both the pilot tone and CDD algorithms were tested. Note that the pilot tone algorithm converges to the optimum weights in all cases, while the decision-directed algorithm only succeeds while the desired laser remains between the two neighboring channels. When the offset gets close to 2 HWHM, thedecision-directed algorithm fails and converges to channel 3, it also converges

to channel three when the desired channel moves halfway between channels 5 and 6. Note that the pilot tone algorithm converges successfully in these cases[3].

Simulations of the CDD and pilot tone LMS algorithms were run for the case of a nonlinear beating term. A seven channel network is simulated with 4 as the desired channel. Channel 2 drifts in frequency and collides with channel 3 thus creating a nonlinear beating signal. For the pilot tone algorithm, the worst case was assumed and there was no nonlinear energy present in the pilot tone spectrum so the algorithm is "blind" to the existence of the beating term. Figure 28 shows the linear and nonlinear crosstalk power present in the filter output during filter convergence. Figure 28a and Figure 28b show the output for the CDD and pilot tone algorithms, respectively. Note both algorithms cancel both types of crosstalk. Their success is due to the fact that the light intensity profiles of the two colliding channels and the nonlinear signal are nearly identical. So a single degree of freedom in weight space can cancel all three signals.

---

[3] The reader may note that the optimum SCNR is lower when the desired channel is at three HWHM than when it is at one HWHM, even though both positions are separated from channel five by one HWHM. This effect is due to the influence of channel 6, located at 4 HWHM.

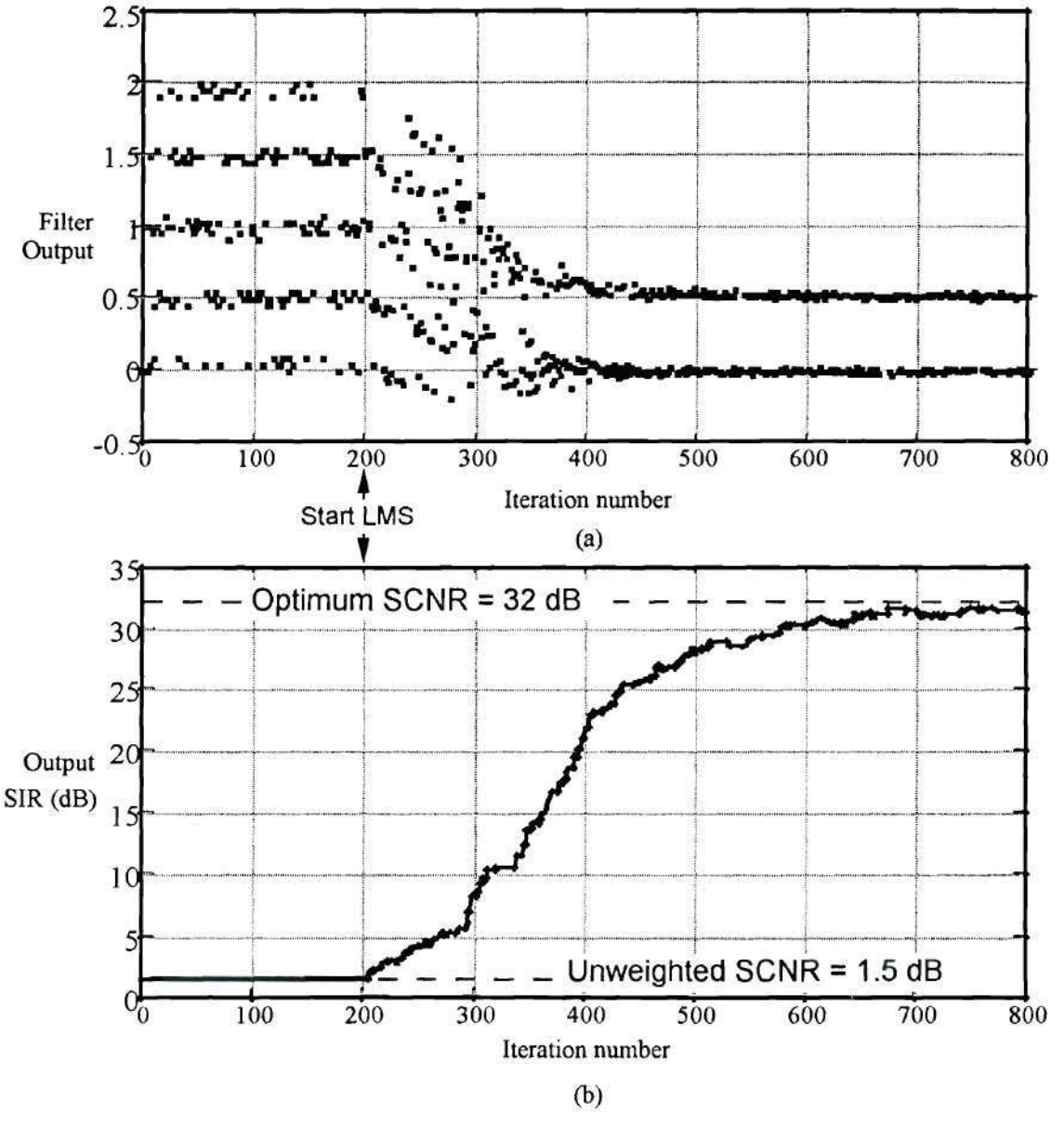


Figure 24: Example of CDD algorithm. Cancellation begins at iteration 200. 19 channels, nine cancellation weights, channel ten is the desired channel, channel and detector spacing is 1.6 HWHM.

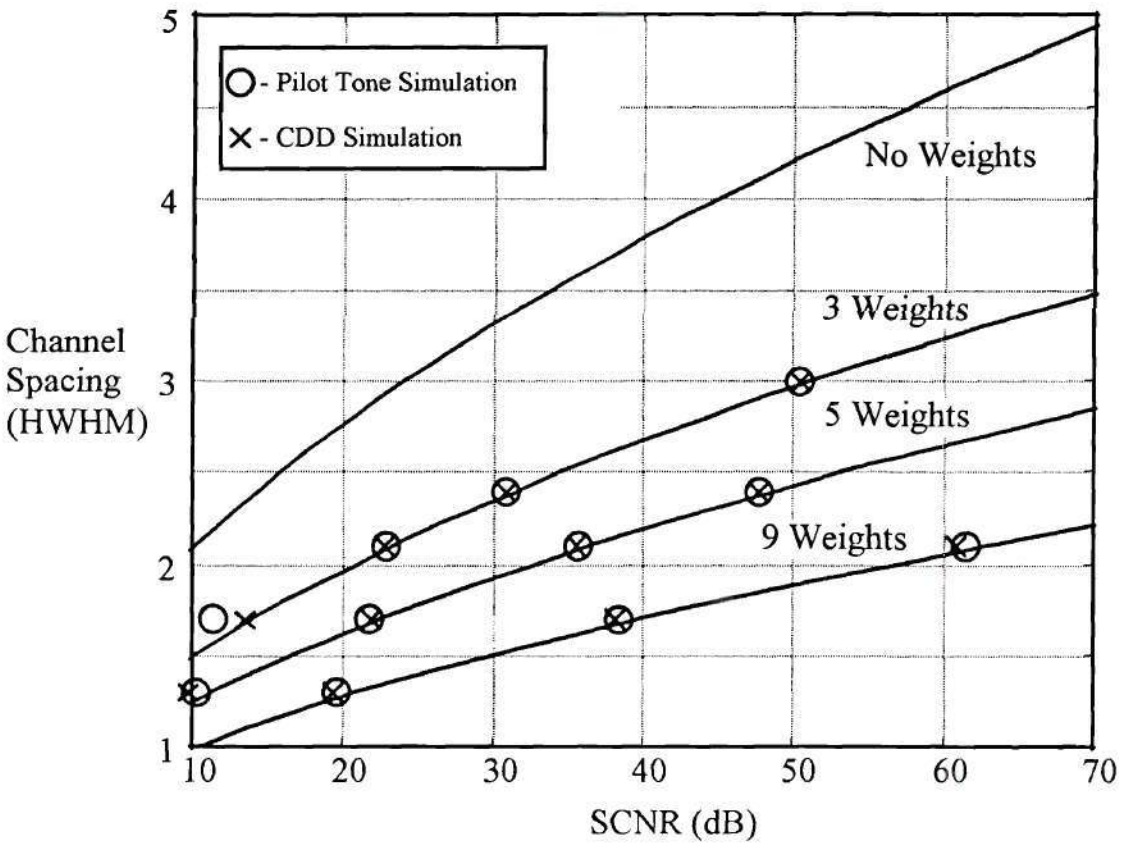


Figure 25: Channel separation required to achieve a given SCNR level for different number of cancellation weights. Curves are upper bounds achieved by optimum weighting, symbols are results achieved by simulation. 19 channels, channel ten the desired.

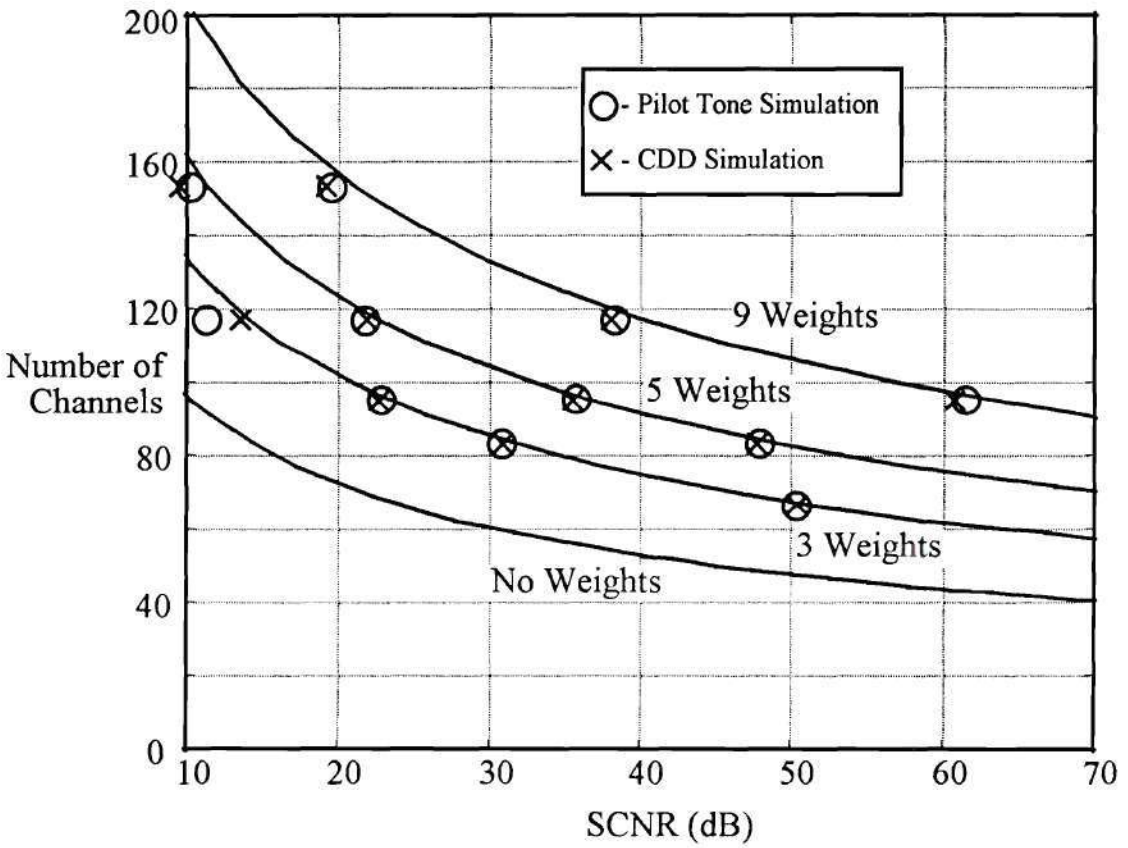


Figure 26: Number of channels that can fit into the 200 nm-wide 1500 nm transmission window as a function of SCNR for different numbers of cancellation weights. Curves are upper bounds achieved by optimum weighting, symbols are results achieved by simulation. 19 channels, channel ten the desired, demultiplexer has one nm resolution.

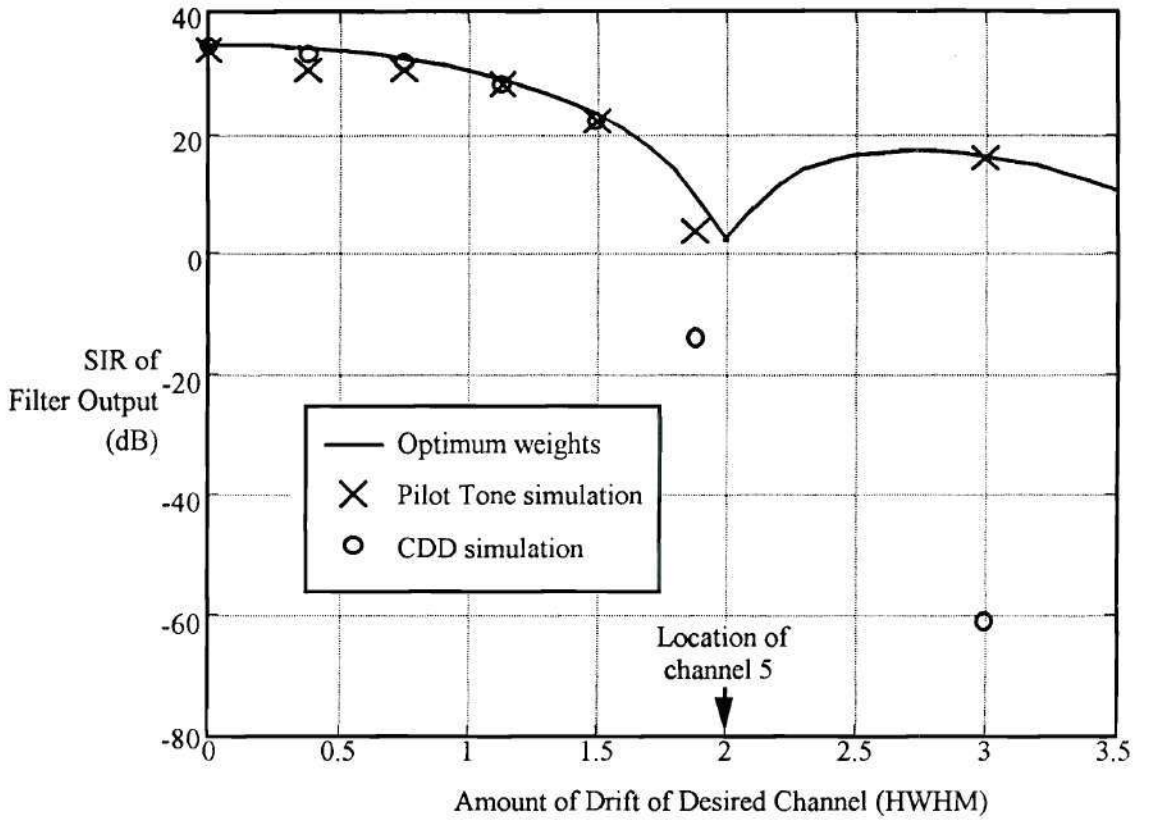
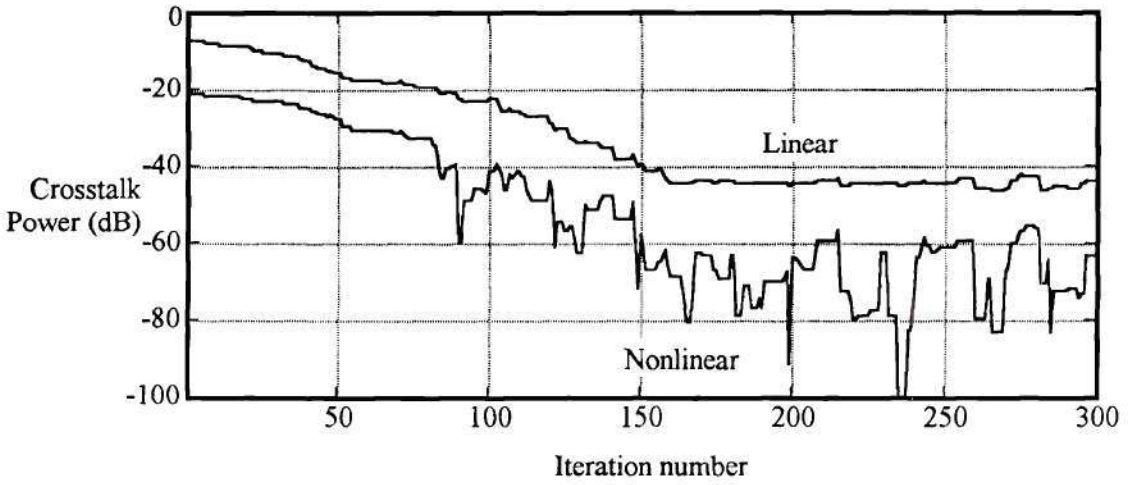
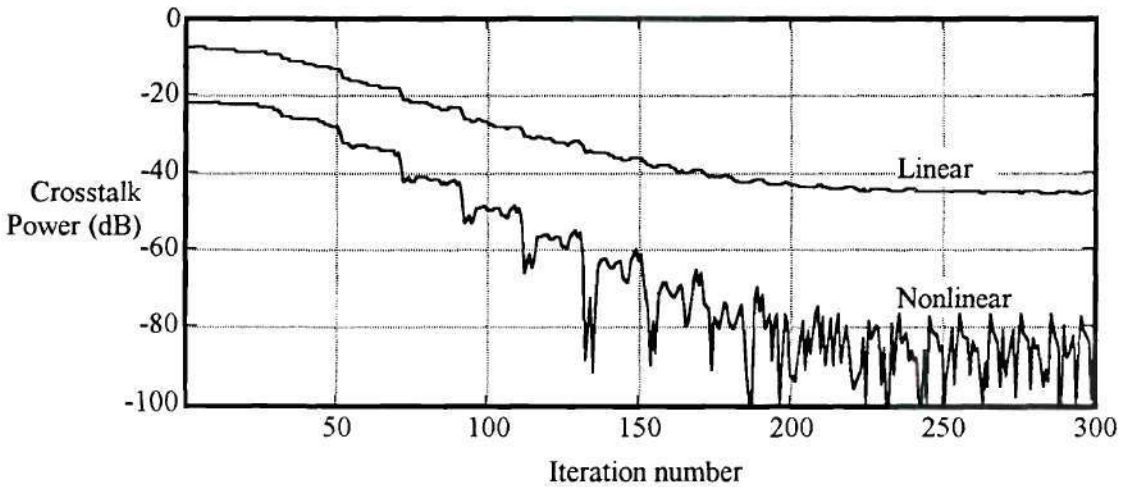


Figure 27: SCNR of canceller output for a laser drifting in frequency. The solid line is the SCNR produced by optimum weights and the symbols are the SCNR produced by simulations of the algorithms. The network has seven channels ad seven detectors with a nominal spacing of two HWHM. Channel four is the desired channel. Note that the Pilot Tone algorithm can accommodate all amounts of drift while CDD fails when channel four drifts near, or past, channel five, located at two HWHM.



(a) Constrained Decision Directed



(b) Pilot Tone

Figure 28: Algorithm performance with nonlinear crosstalk created by beating term between channels two and three. Seven channels and seven detectors spaced at two HWHM. Receiver noise is at -47 dB.

CHAPTER 6.

CONCLUSIONS

We have accomplished several things with this research: We developed a geometrical interpretation of misconvergence in Bussgang type blind equalizers. We exploited this insight to create the Variance Constraint (VC) algorithms, which have robust convergence properties. In the area of WDM fiber-optic receivers we were the first to propose post detection signal processing in order to eliminate crosstalk in WDM receivers. To this end we developed and simulated four algorithms.

We have advanced the understanding of the nature of misconvergence in Bussgang type blind equalizers by linking bad stable points (i.e. points of convergence that have poor residual ISI) to MMSE weights for approximating certain delays. If the channel is known, we find bad stable points by initializing the Bussgang algorithm with the MMSE weights for a delay that produces a low SIR. Because the equalizer problem is inherently underdetermined such delays always exist for nontrivial channels. Specifically, if the first element of the channel impulse response has the greatest magnitude, then the extremely long delays have a poor SIR. Similarly, if the final element has the largest magnitude, then extremely short delays produce poor SIR. Finally, if the largest magnitude of the impulse response is anywhere in the middle, both short and long delays will be poor. We performed a Monte-Carlo simulation of the decision directed and constant modulus algorithms and successfully found bad stable points as predicted.

We have developed the VC algorithms, a new class of blind equalizers that achieve more robust convergence properties than standard Bussgang algorithms by combining new cost functions with a linear constraint. The cost function for the p^{th} variance constraint algorithm VC_p is $\sigma^2_{|y|^p}$, the variance of the output modulus raised to the p^{th} power. These

cost functions are related to the well known Godard cost functions. The difference is that the constants in the Godard cost functions are replaced with the expected value of the output magnitude raised to a power. Godard cost functions have local minima near the MMSE delay weights. The VC cost functions, instead, have a single global minimum at the origin and have "valleys" or "troughs" running radially from the origin through the MMSE points. The constrained cost function minima occur where the constrained affine subspace intersects the troughs. A good constraint will intersect the well performing troughs for central delays near the origin and intersects the poor SIR, extreme delay troughs far away from the origin. We suggest $\mathbf{I}_{N/2}$ as both the constraint and the initial weight vector. This constraint pushes stable points for extremely long and short delays farther away because the weights for these delays have their significant weights bunched in the first or last elements and are nearly orthogonal to $\mathbf{I}_{N/2}$. Therefore, if a weight vector approximates an extreme delay and satisfies the constraint $\mathbf{I}_{N/2}^{T}\mathbf{W}=1$ it must have a large magnitude. If the equalizer is long enough, for example, greater than twice the length of the channel impulse response ($N>2M$), delays of around $(N+M)/2$ give good performance in all cases. Their MMSE weights are closest to the constraint and initial weights of $\mathbf{I}_{N/2}$. We used a Monte-Carlo simulation of the VC_1 and VC_2 algorithms with the $\mathbf{I}_{N/2}$ constraint to demonstrate their robust convergence properties.

We believe that our VC algorithms offer advantages over other constraint-based algorithms presented recently in the literature. [Kamel '94] requires a minimum gain to work, [Verdu '93] needs an additional allpass filter to complete the equalization and [Vembu '94] requires that high exponentiation of the channel output be fed back into the algorithm. The VC algorithms have none of these requirements.

The problem of crosstalk cancellation in WDM fiber-optic receivers has many similarities to the problem of equalization. Some WDM demultiplexers make available an

array of photo-detector outputs that can be weighted and combined to select a desired channel and eliminate crosstalk from other channels. We are the first to propose postdetection signal processing of this array of photodetectors for eliminating crosstalk in WDM receivers. This approach should work with any broadcast-and-select WDM network. But it also applies to any case where crosstalk from signals with different optical wavelengths is a problem, for example, residual signal left in a fiber after passing through a drop/add node in a network. This technique cannot remove crosstalk from signals with the same wavelength as the desired signal. Our simulations show that this simple signal processing technique in the receiver can more than double the usable capacity of a WDM optical link.

We developed four algorithms for crosstalk cancellation. Two algorithms, decision directed and constant modulus, were taken from the field of blind equalization and array processing. They were augmented with peak forcing, a weight management technique related to tap centering, to ensure that the filter selects the desired channel. Both algorithms require that the d.c. component of the detector currents be filtered out. We performed a Monte-Carlo simulation with over 1000 channels for both algorithms. We successfully converged to the desired channel every time. Two new algorithms were developed. The constrained decision-directed algorithm cancels crosstalk between OOK signals or any signal where the absence of light is one of the symbols. We showed that CDD relies on the same cost function as VC_1 (modified for a $\{0,1\}$ signal). However, CDD does not require estimation of the channel output statistics. Like the VC algorithms, CDD uses a linear constraint on the weights. Choice of constraint requires knowledge of the nominal wavelength of the desired laser signal. Specifically, we must know which detector receives the most energy from the desired signal. The final algorithm is the pilot-tone algorithm. It requires the addition of simple sinusoidal signals to the transmitted data.

The sinusoids are used as training signals for an LMS algorithm. It has an advantage over standard training sequence LMS because data transmission and training occur simultaneously. In other words, data transmission and weight training are multiplexed in frequency instead of in time. Because the algorithm is not blind it works with any type of intensity-modulated data, does not require any knowledge of the laser wavelengths, and has a greater ability to follow a drifting laser.

6.1 Future work

We would like to study several aspects of the VC algorithms. Among them are: extend the analysis of the VC algorithms to multilevel PAM and QAM constellations, extend the analysis to include AWGN, and compare the algorithms to Bussgang algorithms augmented with tap centering.

Both the Sato algorithm and CMA work with PAM and QAM constellations and we believe that the corresponding VC algorithms, VC_1 and VC_2, will also. We want to consider AWGN because it may alter the locations of the constrained cost function minima. Tap centering has been successful in simulations in the literature [Foschini '85, Ding '95] and peak forcing worked well for crosstalk cancellation in our research. We would like to compare and contrast this technique with the VC algorithms in more detail.

The VC_1 algorithm shows special promise because it gives an algorithm with robust convergence and with the final convergence properties similar to DDA. However, VC_1 is more complex than the DDA because it requires the estimation of $E\{|y|\}$. We want to study how well we must estimate $E\{|y|\}$ and how it effects the convergence speed of VC_1. If convergence time is quick enough, it may eliminate the need to switch to DDA once the eye is opened. We will also search for ways to modify all of the VC algorithms to eliminate the estimation altogether.

We believe that eigenanalysis techniques like the MUSIC algorithm, described in Chapter 2, may be modified for WDM crosstalk cancellation. If this could be done, we would have an algorithm that is blind, like CDD, but is data-format free like the pilot tone algorithm.

Finally, for all of the algorithms presented in this thesis, we would like to extend the work beyond computer simulations. We suggest a three step approach: first, use actual data recorded from receivers to exercise the algorithm simulations, secondly, feed the receiver outputs directly into the computer for real-time simulations at realistic data rates, and finally, build prototype receivers with the algorithms built in.

APPENDIX A

DERIVATION OF THIRD TERM OF BUSSGANG POWER SERIES EXPANSION (EQUATION 37)

This appendix will evaluate the third term of the power series expansion of the Bussgang nonlinear error function:

$$\mathbf{t} = E\{\frac{e''(\mathrm{a_{m\text{-}d}})}{2}\mathbf{H}\mathbf{a}_m\mathbf{I}_d^T(\mathbf{H}^{\#}\mathbf{H}\text{-}\mathbf{I})\mathbf{a}_m\mathbf{a}_m^T(\mathbf{H}^{\#}\mathbf{H}\text{-}\mathbf{I})\mathbf{I}_d\}. \quad (65)$$

First note the $e(y)$ is an odd function so $e''(y)$ is also odd and $e''(a_{m\text{-}d})=e''(1)a_{m\text{-}d}$. This gives:

$$\mathbf{t} = \frac{e''(1)}{2}\mathbf{H}E\{\mathbf{a}_m\mathbf{I}_d^T\mathbf{P}_\perp\mathbf{a}_m\mathbf{a}_m^T\mathbf{P}_\perp\mathbf{I}_d a_{m\text{-}d}\} \quad (66)$$

where $\mathbf{P}_\perp$ is the projection operator into the null space of $\mathbf{H}$.

Let $\mathbf{E}$ be expected value portion of Equation (66), then the i^{th} element of $\mathbf{E}$ is:

$$E_i = E\left\{a_{m-d}a_{m-i}\mathbf{I}_d^T\mathbf{P}_\perp\mathbf{a}_m\mathbf{a}_m^T\mathbf{P}_\perp\mathbf{I}_d\right\} = E\left\{\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}p_{jd}p_{dk}a_{m-j}a_{m-k}a_{m-i}a_{m-d}\right\}, \quad (67)$$

where p_{ij} is an element of $\mathbf{P}_\perp$. Since d is fixed for all elements and i is fixed within each element the terms in the summation are nonzero for only three cases: ($d=i$ and $j=k$), ($d\neq i$, $k=i$ and $j=d$) and ($d\neq i$, $k=d$ and $j=i$). So if $d\neq i$:

$$E_i = 2p_{dd}p_{id}, \tag{68}$$

and if $d=i$:

$$E_d = \sum_{k=0}^{N} p_{kd}^2 = \left\|\mathbf{P}_{\perp,d}\right\|^2 = p_{dd}, \tag{69}$$

where we use two properties of projection matrices: $\mathbf{P}$ is symmetric and, $\mathbf{P}^2=\mathbf{P}$. Equation (68) and (69) can be recombined to give back the vector equation:

$$\mathbf{E} = 2p_{dd}\mathbf{P}_{\perp,d} + \left(p_{dd} - 2p_{dd}^2\right)\mathbf{I}_d. \tag{70}$$

Now return to $\mathbf{t}$

$$\begin{aligned}
\mathbf{t} &= e''(1)\mathbf{HE} \\
&= 2e''(1)p_{dd}\mathbf{HP}_{\perp,d} + e''(1)\left(p_{dd} - 2p_{dd}^2\right)\mathbf{HI}_d \\
&= \mathbf{0} + e''(1)\left(p_{dd} - 2p_{dd}^2\right)\mathbf{H}_d \\
&= e''(1)\left(p_{dd} - 2p_{dd}^2\right)\mathbf{H}_d
\end{aligned}, \tag{71}$$

where $\mathbf{P}_{\perp,d}$ and $\mathbf{H}_d$ are the d^{th} column of $\mathbf{P}_\perp$ and $\mathbf{H}$ respectively. The first term is $\mathbf{0}$ because, by definition, $\mathbf{HP}_\perp=\mathbf{0}$.

We now relate Equation (71) to SIR_d, the SIR produced by the MMSE weights for a delay of d samples. First define θ_d as the angle between the space spanned by $\mathbf{H}$ and $\mathbf{I}_d$. Since $p_{dd} = \sin^2(\theta_d)$ and $\text{SIR}_d = \dfrac{\cos^2(\theta_d)}{\sin^2(\theta_d)}$ then $p_{dd} = \dfrac{1}{\text{SIR}_d + 1}$ giving:

$$\mathbf{t} = \frac{e''(1)}{2}\left(\frac{1}{\mathrm{SIR}_d + 1} + \frac{2}{(\mathrm{SIR}_d + 1)^2}\right)\mathbf{H}_d \approx \frac{e''(1)}{2\mathrm{SIR}_d}\mathbf{H}_d, \tag{72}$$

where the approximation holds for high SIR_d.

APPENDIX B

EXISTANCE OF BAD LOCAL MINIMA IN THE DDA COST FUNCTION

Theorem: The DDA cost function has local minima at all points with weights such that $\mathbf{T}=\mathbf{G}^T\mathbf{W}$ has only n nonzero values, n an odd number, and all n elements are equal to $\pm\binom{n-1}{(n-1)/2}2^{-(n-1)}$.

First we prove that if two vectors differ by only the arithmetic signs of their elements their cost functions are identical. This allows us to only prove the theorem for the case where all of the nonzero elements are positive.

Lemma: Let $J(\mathbf{T})=E\{f(\mathbf{T}^T\mathbf{S})\}$. If $\mathbf{T}_1$ and $\mathbf{T}_2$ are identical except that their elements may differ in sign, i.e., $|t_{1,i}|=|t_{2,i}|$ for all i. Then $J(\mathbf{T}_1)=J(\mathbf{T}_2)$.

Proof: Let $\mathbf{T}_1$ and $\mathbf{T}_2$ have length N and be identical except for the arithmetic sign of the k^{th} element t_k,. (i.e., $t_{1,k}=-t_{2,k}$), then

$$J(\mathbf{T}_1)=E\left\{f\left(\mathbf{T}_1^T\mathbf{S}\right)\right\}=2^{-N}\sum_{\underline{\underline{\mathbf{S}}}}f\left(\mathbf{T}_1^T\mathbf{S}\right)=2^{-N}\sum_{\underline{\underline{\mathbf{S}}}}f\left(\sum_{i=1}^{N}t_{1,i}s_i\right). \tag{73}$$

Where $\underline{\underline{\mathbf{S}}}$ is the set of all possible outcomes of $\mathbf{S}$ ($\mathbf{S}$ is a vector of i.i.d. symbols with a $\{-1,1\}$ distribution). Now let $\underline{\underline{\mathbf{S}}}^+$ be the set of all possible $\mathbf{S}$ where $s_k=+1$ and let $\underline{\underline{\mathbf{S}}}^-$ be the set of all $\mathbf{S}$ such that $s_k=-1$. Clearly $\underline{\underline{\mathbf{S}}}^+\cup\underline{\underline{\mathbf{S}}}^-=\underline{\underline{\mathbf{S}}}$ and $\underline{\underline{\mathbf{S}}}^+\cap\underline{\underline{\mathbf{S}}}^-=\phi$. Now returning to Equation (73) we have:

$$
\begin{aligned}
J(\mathbf{T}_1) &= 2^{-N}\sum_{\underline{\mathbf{S}}^+} f\left(t_{1,k}+\sum_{i\neq k} t_{1,i}s_i\right) + 2^{-N}\sum_{\underline{\mathbf{S}}^-} f\left(-t_{1,k}+\sum_{i\neq k} t_{1,i}s_i\right) \\
&= 2^{-N}\sum_{\underline{\mathbf{S}}^+} f\left(-t_{2,k}+\sum_{i\neq k} t_{1,i}s_i\right) + 2^{-N}\sum_{\underline{\mathbf{S}}^-} f\left(t_{2,k}+\sum_{i\neq k} t_{1,i}s_i\right) \\
&= 2^{-N}\sum_{\underline{\mathbf{S}}^-} f\left(-t_{2,k}+\sum_{i\neq k} t_{2,i}s_i\right) + 2^{-N}\sum_{\underline{\mathbf{S}}^+} f\left(t_{2,k}+\sum_{i\neq k} t_{2,i}s_i\right) \\
&= 2^{-N}\sum_{\underline{\mathbf{S}}^-} f\left(\mathbf{T}_2^T\mathbf{S}\right) + 2^{-N}\sum_{\underline{\mathbf{S}}^+} f\left(\mathbf{T}_2^T\mathbf{S}\right) = 2^{-N}\sum_{\underline{\mathbf{S}}} f\left(\mathbf{T}_2^T\mathbf{S}\right) = J(\mathbf{T}_2)
\end{aligned}
\quad . \quad (74)
$$

Successive applications of the above will account for any number of sign changes. Q.E.D.

Proof of main theorem: The DDA cost function is:

$$
J(\mathbf{T}) = E\left\{\left(|y|-1\right)^2\right\} = E\left\{\left(\left|\mathbf{T}^T\mathbf{S}\right|-1\right)^2\right\}. \quad (75)
$$

$\mathbf{T}_s$ is a local minima of J if the gradient of J equals $\mathbf{0}$ at $\mathbf{T}_s$. and if the Hessian of J is positive definite (The Hessian matrix, $\mathbf{K}$, is a N-by-N matrix, where $k_{ij} = \frac{\partial^2 J}{\partial_i \partial_j}$). For DDA the gradient is:

$$
\nabla_{\mathbf{T}} J = E\left\{\mathbf{S}\left(\mathbf{T}^T\mathbf{S} - \operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right)\right\} = E\left\{\mathbf{S}\mathbf{T}^T\mathbf{S}\right\} - E\left\{\mathbf{S}\operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = \mathbf{T} - E\left\{\mathbf{S}\operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\}. \quad (76)
$$

Let $\mathbf{T}_n$ have n nonzero elements, n is odd. Let each nonzero element equal $\binom{n-1}{(n-1)/2}2^{-(n-1)}$, which will be labeled c_n. Now evaluate $E\left\{\mathbf{S}\operatorname{sgn}\left(\mathbf{T}_n^T\mathbf{S}\right)\right\}$ for each element of $\mathbf{T}_n$. If $t_{n,i}=0$ then s_i is independent of sgn(y) so $E\left\{s_i \operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = 0$. Next, look

at the case where $t_{n,i}$ is nonzero and equal to c_n. First note that if s_i and y agree in sign $s_i\text{sgn}(y)=+1$, and if they disagree $s_i\text{sgn}(y)=-1$. Therefore, the expected value is equal to the fraction of cases where s_i and y agree in sign minus the fraction of cases where they disagree. There are n nonzero weighted channels. Since, by assumption, they all have weight c_n, we need only count all of the cases where less than $(n\text{-}1)/2$ symbols disagree with s_i, subtract the number of cases that more than $(n\text{-}1)/2$ symbols disagree with s_i and, finally, divide the result by the total number of cases, 2^n.

$$E\left\{s_i \operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = \left(2\sum_{l=0}^{(n-1)/2}\binom{n-1}{l} - 2\sum_{l=\frac{n-1}{2}+1}^{n-1}\binom{n-1}{l}\right)2^{-n}, \tag{77}$$

where the 2 in front of each summation is due to the fact that s_i can equal +1 or -1. Now use the identity

$$\binom{m}{p} = \binom{m}{m-p} \tag{78}$$

to simplify Equation (77) giving

$$E\left\{s_i \operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = \binom{n-1}{(n-1)/2}2^{-n-1} = c_n = t_i. \tag{79}$$

Thus each element of the expected value equals $t_{n,i}$. and the entire vector becomes

$$E\left\{\mathbf{S}\operatorname{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = \mathbf{T}. \tag{80}$$

Substituting Equation (80) into Equation (76) gives

$$\nabla_{\mathbf{T}} J = \mathbf{T} - E\left\{\mathbf{S}\,\mathrm{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\} = \mathbf{T} - \mathbf{T} = \mathbf{0}. \tag{81}$$

This proves that $\mathbf{T}_n$ is a critical point of the cost function.

Before we consider the Hessian, we will define an *interior point*.

Definition: A vector $\mathbf{T}$ is an *interior point* if there exist no $\mathbf{S} \in \underline{\underline{\mathbf{S}}}$ such that $\mathbf{T}^T\mathbf{S}=0$ (i.e., it is impossible for the filter output to equal 0 for any input sequence).

If $\mathbf{T}$ is an interior point an open neighborhood exists around $\mathbf{T}$ such that the quantity $E\left\{\mathbf{S}\,\mathrm{sgn}\left(\mathbf{T}^T\mathbf{S}\right)\right\}$ is a constant. In that case the Hessian is simply $\mathbf{I}$, which is positive definite. All $\mathbf{T}_n$ are interior points because the output can only equal kc_n, $k=\{-n, -n+2, \ldots, -1, 1, \ldots, n-2, n\}$ and cannot equal zero. Q.E.D.

APPENDIX C

ANALYSIS OF PILOT TONE STEADY STATE WEIGHTS

This appendix has three parts. First we examine $\mathbf{W}_{MSE}$ as a function of $\mathbf{R}_{SS}$ for a noiseless overdetermined receiver and an underdetermined receiver. Second we reintroduce receiver noise and show that the MSE weights now contain a perturbation that depends on $\mathbf{R}_{SS}$. Finally we develop an approximation for the excess MSE that is generated when weights developed using pilot tones are applied to the data signals.

Overdetermined: With the receiver noise neglected, the covariance matrix, $\mathbf{R}_{ZZ}$, becomes $\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T$. It is singular and therefore its inverse does not exist so the Wiener-Hopf equation must be solved using the pseudo-inverse [Scharf '91].

Consider the general linear relation, $\mathbf{A}\mathbf{x}=\mathbf{b}$, where the objective is to solve for $\mathbf{x}$. If $\mathbf{A}$ is singular then there may be no solution. However there always exists an $\mathbf{x}$ that is best in the least squares sense, the $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is as close to $\mathbf{b}$ as possible, the $\mathbf{x}$ that minimizes the value $||\mathbf{A}\mathbf{x}-\mathbf{b}||$. Although this $\mathbf{x}$ may not be unique (there may be an infinite number), the added restriction that the norm of $\mathbf{x}$ must be as small as possible will make $\mathbf{x}$ unique. The Penrose pseudo-inverse, denoted as $\mathbf{A}^{\#}$, solves for this $\mathbf{x}$,

$$\mathbf{x} = \mathbf{A}^{\#}\mathbf{b}. \tag{82}$$

Note that the pseudo-inverse exists for nonsquare matrices. In the case of an n by m matrix the pseudo-inverse will be m by n.

The Penrose pseudo-inverse is described in many linear algebra texts, some fundamental properties that we need are listed here:

1. $\mathbf{A}^{\#} = (\mathbf{A}\mathbf{A}^{T})^{-1}\mathbf{A}$, if $\mathbf{A}\mathbf{A}^{T}$ is nonsingular

2. $\mathbf{A}^{\#}\mathbf{A} = \mathbf{P}_{\mathbf{A}}$, $\mathbf{P}_{\mathbf{A}}$ is the projection operator onto the space spanned by the rows of $\mathbf{A}$

3. $\mathbf{A}\mathbf{A}^{\#} = \mathbf{P}_{\mathbf{A}^{T}}$, $\mathbf{P}_{\mathbf{A}^{T}}$ is the projection onto the space spanned by the columns of $\mathbf{A}$

4. $(\mathbf{AB})^{\#}=\mathbf{B}^{\#}\mathbf{A}^{\#}$, if and only if the space spanned by the rows of $\mathbf{A}$ is the same as the space spanned by the columns of $\mathbf{B}$

If either the row or column space has full rank then the corresponding projection operator is the identity matrix $\mathbf{I}$. If $\mathbf{A}$ is flat and has full row rank then $\mathbf{A}\mathbf{A}^{\#} = \mathbf{I}$, if $\mathbf{A}$ is tall and has full column rank then $\mathbf{A}^{\#}\mathbf{A} = \mathbf{I}$.

Using pseudo-inverses, the solution to the Wiener-Hopf equation for the overdetermined noise-free case becomes

$$\mathbf{W}_{\mathrm{MSE}} = (\mathbf{G}\mathbf{R}_{\mathrm{SS}}\mathbf{G}^{T})^{\#}\mathbf{G}\mathbf{R}_{\mathrm{SS}}. \tag{83}$$

If $\mathbf{R}_{\mathrm{SS}}$ is invertible and $\mathbf{G}$ has full row rank (true if all lasers are transmitting at different frequencies) then the conditions for Property 4 hold and $(\mathbf{G}\mathbf{R}_{\mathrm{SS}}\mathbf{G}^{T})^{\#} = (\mathbf{G}^{T})^{\#}\mathbf{R}_{\mathrm{SS}}^{-1}\mathbf{G}^{\#}$ so

$$\mathbf{W}_{\mathrm{MSE}} = (\mathbf{G}^{T})^{\#}\mathbf{R}_{\mathrm{SS}}^{-1}\mathbf{G}^{\#}\mathbf{G}\mathbf{R}_{\mathrm{SS}}. \tag{84}$$

When $\mathbf{G}$ is tall, the quantity $\mathbf{G}^{\#}\mathbf{G} = \mathbf{I}$, so Equation (84) is rewritten as

$$\mathbf{W}_{MSE} = (\mathbf{G}^T)^{\#}. \tag{85}$$

This is the same result as in the invertible case except that the pseudo-inverse replaces the actual inverse. Again note that the optimum weights do not depend on $\mathbf{R}_{SS}$ so the pilot tone weights would be optimum for the data also.

Underdetermined: In this case, $\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T$ is invertible so the Wiener-Hopf solution becomes

$$\mathbf{W}_{MSE} = (\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{R}_{SS}. \tag{86}$$

Property 4 does not hold in this case but we can still postmultiply both sides of this equation by the term $\mathbf{G}^T(\mathbf{G}^T)^{\#}$. However when $\mathbf{G}$ is flat this is not $\mathbf{I}$ but $\mathbf{P}_{\mathbf{G}}$, the projection operator onto the space spanned by the rows of $\mathbf{G}$. The result is

$$\mathbf{W}_{MSE}\mathbf{P}_{\mathbf{G}} = (\mathbf{G}^T)^{\#}. \tag{87}$$

So the optimum weights do vary with $\mathbf{R}_{SS}$, however the projection of the weights into the space spanned by the rows of $\mathbf{G}$ is independent of $\mathbf{R}_{SS}$ and is equal to $(\mathbf{G}^T)^{\#}$.

We will now reintroduce receiver noise. The first step is to find an expression for $\mathbf{R}_{ZZ}^{-1}$, the inverse of the covariance matrix of the detector currents. With receiver thermal noise added, the covariance matrix has the form $\mathbf{R}_{ZZ} = \mathbf{R} + \sigma^2\mathbf{I}$ where $\mathbf{R} = \mathbf{G}\mathbf{R}_{SS}\mathbf{G}^T$. $\mathbf{R}$ is symmetric ($\mathbf{R} = \mathbf{R}^T$) and positive semi-definite ($\mathbf{W}^T\mathbf{R}\mathbf{W} \geq 0$ for all $\mathbf{W}$ not equal to 0) so all of its eigenvalues are greater than or equal to zero and all of the eigenvectors are mutually orthogonal. Additionally, for the underdetermined and uniquely determined networks, $\mathbf{R}$ is nonsingular and has a complete set of strictly positive eigenvalues. $\mathbf{R}$ can then be

diagonalized by an orthonormal matrix **Q** (all columns are mutually orthogonal with unit magnitude and $\mathbf{Q}^{-1} = \mathbf{Q}^{T}$)

$$\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^{T}. \tag{88}$$

The columns of **Q** are the eigenvectors of **R** and Λ is a diagonal matrix with the diagonal elements equal to the corresponding eigenvalues. $\mathbf{R}_{ZZ}$ is now

$$\mathbf{R}_{ZZ} = \mathbf{Q}(\Lambda + \sigma^2\mathbf{I})\mathbf{Q}^{T}. \tag{89}$$

We now see that $\mathbf{R}_{ZZ}$ has the same eigenvectors as **R** and its eigenvalues have increased by the amount σ^2. So receiver noise does not change the eigenvectors, only the eigenvalues. The inverse of $\mathbf{R}_{ZZ}$ is

$$\mathbf{R}_{ZZ}^{-1} = \mathbf{Q}(\Lambda + \sigma^2\mathbf{I})^{-1}\mathbf{Q}^{T}, \tag{90}$$

where the i^{th} diagonal element of $(\Lambda + \sigma^2\mathbf{I})^{-1}$ is $1/(\lambda_i + \sigma^2)$. Now if all of the eigenvalues of **R** are greater than σ^2, the diagonal terms can be expanded into a power series,

$$\frac{1}{\lambda_i + \sigma^2} = \sum_{n=0}^{\infty} (-1)^n \frac{\sigma^{2n}}{\lambda_i^{n+1}}. \tag{91}$$

Using these expansions, $\mathbf{R}_{ZZ}^{-1}$ can also be expanded as

$$\mathbf{R}_{ZZ}^{-1} = \mathbf{R}^{-1} - \sigma^2\mathbf{R}^{-2} + \sigma^4\mathbf{R}^{-3} - \sigma^6\mathbf{R}^{-4} + \ldots = \sum_{n=0}^{\infty} (-1)^n\sigma^{2n}\mathbf{R}^{-(n+1)} \quad (92a)$$

$$= (\mathbf{I} - \sigma^2\mathbf{R}_{ZZ}^{-1})\mathbf{R}^{-1}. \quad (92b)$$

For the overdetermined case some of the eigenvalues of $\mathbf{R}$ are zero. The spaced spanned by $\mathbf{R}_{ZZ}$ can be broken into two subspaces, the first subspace is spanned by the columns of $\mathbf{R}$ and is called the signal subspace; it is the same space that is spanned by the columns of $\mathbf{G}$. The second subspace is the null space of $\mathbf{R}$ and is called the noise subspace. The signal subspace is spanned by the eigenvectors associated with the nonzero eigenvalues of $\mathbf{R}$, they are also eigenvectors of $\mathbf{R}_{ZZ}$ with eigenvalues $\lambda_i+\sigma^2$. The noise subspace is spanned by the remaining eigenvectors of $\mathbf{R}_{ZZ}$ associated with the eigenvalues of σ^2. Equation (90) can be rewritten as

$$\mathbf{R}_{ZZ}^{-1} = \mathbf{Q}_n\mathbf{Q}_n^{T}/\sigma^2 + \mathbf{Q}_S\Lambda_S^{-1}\mathbf{Q}_S^{T}, \quad (93)$$

where the columns of $\mathbf{Q}_S$ are the eigenvectors in the signal subspace, and the columns of $\mathbf{Q}_n$ are the eigenvectors in the noise subspace. The matrix Λ_S is square and diagonal and the i^{th} diagonal element is the eigenvalue $\lambda_i+\sigma^2$.

The second term of Equation (93) can be expanded using the same power series as in the previous analysis producing

$$\mathbf{R}_{ZZ}^{-1} = \mathbf{Q}_n\mathbf{Q}_n^{T}/\sigma^2 + \sum_{n=0}^{\infty} (-1)^n\sigma^{2n}\mathbf{R}^{\#(n+1)}$$

$$= \mathbf{Q}_n\mathbf{Q}_n^{T}/\sigma^2 + \mathbf{R}^{\#} - \sigma^2\mathbf{R}^{\#2} + \sigma^4\mathbf{R}^{\#3} - \sigma^6\mathbf{R}^{\#4} + \ldots, \quad (94)$$

which is similar to the result in (92a) except for an additional term due to the noise subspace.

With $\mathbf{R}_{ZZ}^{-1}$ determined, we are now prepared to return to the Wiener-Hopf equation and look at the optimum weights with noise added. For the uniquely determined receiver we get

$$\begin{aligned}\mathbf{W}_{MSE} &= \mathbf{R}_{ZZ}^{-1}\mathbf{G}\mathbf{R}_{SS} = \mathbf{R}_{ZZ}^{-1}\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^{T}(\mathbf{G}^{T})^{-1} \\ &= (\mathbf{I}-\sigma^2\mathbf{R}_{ZZ}^{-1})\mathbf{R}^{-1}\mathbf{R}(\mathbf{G}^{T})^{-1} = (\mathbf{G}^{T})^{-1}-\sigma^2\mathbf{R}_{ZZ}^{-1}(\mathbf{G}^{T})^{-1}, \end{aligned} \tag{95}$$

and for the overdetermined case we get

$$\begin{aligned}\mathbf{W}_{MSE} &= (\mathbf{Q}_n\mathbf{Q}_n^{T})\mathbf{G}\mathbf{R}_{SS}/\sigma^2 + (\mathbf{G}^{T})^{\#} - \sigma^2\mathbf{R}_{ZZ}^{-1}(\mathbf{G}^{T})^{\#} \\ &= (\mathbf{G}^{T})^{\#} - \sigma^2\mathbf{R}_{ZZ}^{-1}(\mathbf{G}^{T})^{\#}. \end{aligned} \tag{96}$$

The first term in the above is zero because $\mathbf{Q}_n\mathbf{Q}_n^{T}$ is the projection operator into the noise subspace that, by definition, is orthogonal to all of the columns of $\mathbf{G}$, so $(\mathbf{Q}_n\mathbf{Q}_n^{T})\mathbf{G} = \mathbf{0}$. For the underdetermined case, we can postmultiply $\mathbf{W}_{MSE}$ by $\mathbf{G}^{T}(\mathbf{G}^{T})^{\#} = \mathbf{P}_{\mathbf{G}}$ to give

$$\mathbf{W}_{MSE}\mathbf{P}_{\mathbf{G}} = \mathbf{R}_{ZZ}^{-1}\mathbf{G}\mathbf{R}_{SS}\mathbf{G}^{T}(\mathbf{G}^{T})^{\#} = (\mathbf{G}^{T})^{\#} - \sigma^2\mathbf{R}_{ZZ}^{-1}(\mathbf{G}^{T})^{\#}. \tag{97}$$

For each case, the optimum weights, or its projection, consists of an invariant term that does not depend on $\mathbf{R}_{SS}$ plus a term that does depend on $\mathbf{R}_{SS}$ that is scaled by the noise power.

The final step is to examine the excess MSE produced by using $\mathbf{W}_p$ instead $\mathbf{W}_{MSE}$. To do so we need $\Delta\mathbf{W}$, the difference between $\mathbf{W}_p$ and $\mathbf{W}_{MSE}$

$$\Delta\mathbf{W} = \mathbf{W}_p - \mathbf{W}_{MSE} = (\sigma_t^2\mathbf{R}_{zzt}^{-1} - \sigma_p^2\mathbf{R}_{zzp}^{-1})(\mathbf{G}^T)^{\#}, \tag{98}$$

where the subscripts t and p signify training-sequence and pilot tones respectively. The excess mean square error, ΔJ, is the quantity $\mathrm{Tr}(\Delta\mathbf{W}^T\mathbf{R}_{zzt}\Delta\mathbf{W})$ [Treichler '87 pp. 65], where Tr(•) is the trace operator (the sum of the elements of the main diagonal of the matrix).

$$\Delta J=\mathrm{Tr}(\Delta\mathbf{W}^T\mathbf{W}_{zzt}\Delta\mathbf{W})=\mathrm{Tr}[\mathbf{G}^{\#}(\sigma_t^4\mathbf{R}_{zzt}^{-1}+\sigma_p^4(\mathbf{R}_{zzp}\mathbf{R}_{zzt}\mathbf{R}_{zzp})^{-1}-2\sigma_t^2\sigma_p^2\mathbf{R}_{zzp}^{-1})(\mathbf{G}^T)^{\#}]. \tag{99}$$

If we expand $\mathbf{R}_{zzt}^{-1}$ and $\mathbf{R}_{zzp}^{-1}$ using Equation (92a) and take only the first term we get

$$\Delta J \approx \sigma_t^4\mathrm{Tr}[(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{R}_{ss}^{-1}(\mathbf{I}-(\sigma_p/\sigma_t)^2\mathbf{R}_{ss}\mathbf{R}_{pp}^{-1})^2(\mathbf{G}\mathbf{G}^T)^{-1}]. \tag{100}$$

REFERENCES

Agee, B. G.: Convergent Behavior of Modulus-Restoring Adaptive Arrays in Gaussian Interference Environments," *Proceedings of Asilomar Conference on Signals, Systems and Computers*, pp. 818-822, 1988.

Aisawa, S. et al.: "Neural-Processing-Type Optical WDM Demultiplexer," *Journal of Lightwave Technology*, Vol. 11, No. 12, pp. 2130-2139, Dec. 1993.

Batra, A. and Barry, J.: "Blind Cancellation of Co-Channel Interference," *IEEE 1995 Global Telecommunications Conference*, Singapore, vol. 1, pp. 157-162, November 13-17, 1995.

Bellini, S.: "Bussgang Techniques for Blind Equalization," *Proceedings of IEEE-Globecom '86*, pp. 46.1.1-46.1.7, 1986.

Bellini, S.: "Blind Equalization," *Alta Frequenza,* Vol. LVII-N.7, pp. 445-450, September 1988.

Benveniste, A., Goursat, M., and Ruget, G.: "Robust Identification of a Nonminimum Phase System: Blind Adjustment of a Linear Equalizer in Data Communications," *IEEE Transactions on Automatic Control,* Vol. AC-25, No. 3, pp. 385-398, June 1980.

Benveniste, A. and Goursat, M.: "Blind Equalizers," *IEEE Transactions on Communications,* Vol. COM-32, pp. 871-883, August 1984.

Brackett, C.A., Acampora, A.S. et al.: "A Scalable Multiwavelength Multihop Optical Network: A proposal for Research on All-Optical Networks," *Journal of Lightwave Technology*, Vol. 11, No. 5-6, pp. 736-753, May-June 1993.

Bussgang, J.: "Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals," *MIT. Technical Report,* No. 216, March 1952.

Chraplyvy, A. R.: "Limitations on lightwave communications imposed by optical-fiber nonlinearities," *Journal of Lightwave Technology*, Vol. 8, No. 10, pp. 1548-1557, Oct. 1990.

Chung, F. R., Salehi, J. A. and Wei, V. K.: "Optical Orthogonal Codes: Design, Analysis and Applications," *IEEE Transactions on Information Theory*, Vol. 35, No. 3, pp. 824-833, August 1989.

Cremer, C. et al.: "Grating spectrograph in InGaAsP/InP for dense wavelength division multiplexing," *Appl. Phys. Letters*, Vol. 59, No. 6, pp. 627-629, August 5, 1991.

Ding, Z. and Johnson, C. R.: "Existing Gap Between Theory and Application of Blind Equalization," *SPIE Adaptive Signal Processing*, Vol. 1565, pp. 154-165, 1991.

Ding Z., Kennedy, R. A., et al.: "Ill-Convergence of Godard Blind Equalizers in Data Communication Systems," *IEEE Transactions on Communications*, Vol. 39, No. 9, pp. 1313-27, September 1991.

Ding Z., Kennedy, R. A., et al.: "Local Convergence of the Sato Blind Equalizer and Generalizations Under Practical Considerations," *IEEE Transactions on Information Theory*, Vol. 39, No. 1, pp. 129-145, January 1993.

Ding, Z.: "Blind Channel Identification and Equalization Using Spectral Correlation Measurements, Part I: Frequency-Domain Analysis," *Cyclostationarity in Communications and Signal Processing*. Gardner, W. A. ed., New York: IEEE Press, pp. 417-436, 1994.

Foschini, G. J.: "Equalizing Without Altering or Detecting Data," *AT&T Technical Journal*, Vol. 64, pp. 1885-1911, October 1985.

Foschini, G. J. and Vannucci, G.: "Using Spread-Spectrum in a High Capacity Fiber-Optic Local Network," *Journal of Lightwave Technology*, Vol. 6, No. 3, pp. 370-379, March 1988.

Frost, O. L.: "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, Vol. 60, No. 8, pp. 926-935, Aug. 1972.

Gardner, W. A.: "A New Method of Channel Identification," *IEEE Transactions on Communications*, Vol. 39, No. 6, pp. 813-817, June 1991.

Gardner, W. A.: "Exploitation of Spectral Redundancy in Cyclostationary Signals," *IEEE Signal Processing Magazine*, Vol. 8, pp. 14-36, April 1991.

Geckeler, S.: *Optical Fiber Transmission Systems*, Artech House, Norwood MA, 1987.

Griffiths, L. J. and Jim, C. W.: "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 1, pp. 369-371, January 1982.

Godard, D. N.: "Self-Recovering Equalization and Carrier Tracking in Two-Dimensional Data Communication Systems," *IEEE Transactions on Communications*, Vol. COM-28, No. 11, pp. 1867-1875, November 1980.

Goldstein, J. S., Holder, E. J. and Ingram M. A.: "Reduced complexity adaptive structures for jam-resistant satellite communications," *Proceedings of the 1993 Military Communications Conference (MILCOM '93)*, Vol. 3, pp. 1033-1037, Oct. 11-14 1993.

Hargreaves, D., Jessop, P. E. and Haykin, S.: "Neural Network Assisted Wavelength Demultiplexer for Fiber Optic Communications," *Proceedings of the World Congress on Neural Networks*, Vol. 1, July 1993.

Hatzinakos, D. and Nikias, C.: "Blind Equalization Using a Tricepstrum-Based Algorithm," *IEEE Transactions on Communications*, Vol. 39, No. 5, pp. 669-681, May 1991.

Haykin, S.: *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, N. J., 1991 (Chapter 20).

Haykin, S. and Steinhardt, A.: *Adaptive Radar Detection and Estimation*, New York, N. Y., Wiley and Sons, 1992.

Hill, A. M. and Payne, D. B.: "Linear crosstalk in wavelength-division-multiplexed optical fiber transmission systems," *Journal of Lightwave Technology*, Vol. 3, No. 3, pp. 643-651, June 1985.

Hirotsu, H. and Brooke, M. A.: "An Analog Neural Network Chip with Random Weight Change Learning Algorithm," *Proceedings of the International Joint Conference on Neural Networks*, Vol. 3, pp. 3031-4, Nagoya Japan, October 1993.

Ho, K. P. and Kahn, J. M.: "Crosstalk Measurement and Reduction in Dense WDM Systems Using Subcarrier Tone Channel Identification and Linear Cancellation," *International Conference on communications*, Seattle WA, pp. 287-291, June 1995.

Hoenig, M. L., Steiglitz, K. and Gonipath, B.: "Multichannel signal processing for data communications in the presence of crosstalk," *IEEE Trans. on Communications*, Vol. 38, No. 4, pp. 551-558, April 1990.

Humblet, P. A. and Hamby, W. M.: "Crosstalk analysis and filter optimization of single- and double-cavity Fabry-Perot filters," *IEEE Journal on Sel. Areas in Comm.*, Vol. 8, No. 6, pp. 1095-1107, August 1990.

Johnson, C. R.: "Admissibility in Blind Adaptive Channel Equalization," *IEEE Control Systems Magazine*, Vol. 11, No. 1, pp. 3-15, January 1991.

Johnson, D. and Dudgeon, D.: *Array Signal Processing Concepts and Techniques*, Prentice-Hall, Englewood Cliffs, N. J., 1993 (Chapter 7).

Kamel, R. E. and Bar-Ness, Y.: "Anchored Constant Modulus Algorithm (ACMA) for Blind Equalization," *IEEE International Conference on Communications,* New Orleans, pp. 1571-1575, May 1994.

Li, Y., Ding, Z.: "Convergence Analysis of Finite Length Blind Adaptive Equalizers," *IEEE Transactions on Signal Processing*, Vol. 43, No. 9, pp. 2120-2129, September 1995.

Minardi, M. J. and Ingram, M. A.: "Adaptive crosstalk cancellation in dense wavelength division optical networks," *Electronic Letters,* Vol. 28, No. 17, pp. 1621-1622, August 13, 1992.

Minardi, M. J., Ingram M. A. and Goldstein, J. S.: "Transform Domain Techniques for Adaptive Crosstalk Cancellation in Dense Wavelength Division Multiplexing Optical Networks," *Proceedings of the 1993 Military Communications Conference (MILCOM '93)*, Vol. 1, pp. 101-105, Oct. 11-14 1993

Minardi, M. J. and Ingram, M. A.: "Adaptive crosstalk cancellation and laser frequency drift compensation in dense WDM networks", *Journal of Lightwave Technology*, Vol. 13, No. 8, pp. 1624-1635, August 1995.

Monzingo. R. A., and Miller, T. W.: *Introduction to Adaptive Arrays,* New York: John Wiley and Sons, 1980.

O'Farrell, T. and Beale, M.: "Code-Division Multiple-Access (CDMA) Techniques in Optical Fiber LAN," *Second IEE National Conference on Telecomm.*, pp. 111-115, 1989.

Pan, R. and Nikias, C. L.: "The Complex Cepstrum of Higher-Order Cumulants and Nonminimum Phase System Identification," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 36, pp. 186-205, February 1988.

Picchi, G. and Prati, G.: "Blind Equalization and Carrier Recovery Using a "Stop-and-Go" Decision-Directed Algorithm," *IEEE Transactions on Communications,* Vol. COM-35, No. 9, pp. 877-887, September 1987.

Poggiolini, P. and Benedetto, S.: "Performance Analysis of Multiple Subcarrier Encoding of Packet Headers in Quasi-all-Optical WDM Networks," *IEEE Photonics Technical Letters*, Vol. 6, pp. 112-114, 1994.

Proakis, J. G.: *Digital Communication, Second Edition*, New York, N. Y.: McGraw-Hill, 1989 (Chapter 4).

Proakis, J. G. and Nikias, C. L.: "Blind Equalization," *SPIE Adaptive Signal Processing,* Vol. 1565, pp.76-87, 1991.

Prucnal, P. R., Santoro, S. and Sehgal, S.: "Ultrafast All-Optical Synchronous Multiple Access Fiber Networks," *IEEE Journal on Sel. Areas of Comm.*, Vol. SAC-4, No. 5, pp. 547-554, May 1986.

Salehi, J. A.: "Code Division Multiple-Access Techniques in Optical Fiber Networks-Part I: Fundamental Principles," *IEEE Transactions on Communications*, Vol. 37, No. 8, pp. 824-833, August 1989.

Salz, J.: "Digital Transmission Over Cross-Coupled Linear Channels," *AT&T Technical Journal*, Vol. 64, No. 6, pp. 1147-1159, July-August 1985.

Sato, Y.: "A Method of Self-Recovering Equalization for Multilevel Amplitude-Modulation Systems," *IEEE Transactions on Communications*, pp. 679-682, June 1975.

Schmidt, R. O. and Franks, R.: "Multiple DF Signal Processing, an Experimental System." *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 3, pp. 281-290, March 1986.

Shalvi, O. and Weinstein, E.: "New Criteria for Blind Deconvolution of Non-Minimum Phase Systems (Channels)," *IEEE Transactions on Information Theory*, Vol. IT-36, pp. 312-321, March 1990.

Scharf, L.: *Statistical Signal Processing*, Addison-Wesley, pp. 49, 1991.

Shynk, J. J. et al.: "A Comparative Performance Study of Several Blind Equalization Algorithms," *SPIE Adaptive Signal Processing*, Vol. 1565, pp. 102-117, 1991.

Soole, J. B., Scherer, H. P. et al.: "Spectrometer on chip: a monolithic WDM component," *Optical Fiber Communications Conference Technical Digest*, 1992, p. 123, February 1992.

Tamura, S., Nakano, S. and Okazaki, K.: "Optical Code-Multiplex Transmission by Gold Sequences," *Journal of Lightwave Technology*, Vol. LT-3, No. 1, pp. 121-127, February 1985.

Tong, L., Xu, G. and Kailath, T.: "A New Approach to Blind Identification and Equalization of Multipath Channels," *Proceedings 25th Asilomar Conference Signals, Systems, Computers*, Pacific Grove, CA., pp. 856-860, November 1991.

Tong, L., Xu, G. and Kailath, T.: "Blind Channel Identification and Equalization Using Spectral Correlation Measurements, Part II: A Time-Domain Approach," *Cyclostationarity in Communications and Signal Processing*. Gardner W. A. ed., New York: IEEE Press, pp. 437-454, 1994.

Treichler, J. R. and Larimore, M. G.: "New Processing Techniques Based on the Constant Modulus Adaptive Algorithm," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, pp. 420-431, April 1985.

Treichler, J. R., Johnson, C. R. Jr. and Larimore, M. G.: *Theory and Design of Adaptive Filters*, New York: John Wiley and Sons, 1987.

Vembu, S., Verdu, S. et al.: "Convex Cost Functions in Blind Equalization," *IEEE Transactions on Signal Processing*, Vol. 42, No. 8, pp. 1952-1960, August 1994.

Verdu, S.: "On the Selection of Memoryless Adaptive Laws for Blind Equalization in Binary Communications," *Proceedings 6th Int. Conf. on Analysis and Optimization of System*, Nice France, pp. 239-249, June 1984.

Verdu, S., Anderson, B. D. and Kennedy, R. A.: "Blind Equalization Without Gain Identification," *IEEE Transactions on Information Theory*, Vol. 39, No. 1, pp. 292-297, January 1993.

Way, D. A. et al.: "Self routing WDM high-capacity SONET ring network, "*Optical Fiber Communications Conference Technical Digest*, pp. 86, February 1992.

Weerackody, V. and Kassam, S. A.: "Blind Adaptive Equalization Using Dual-Mode Algorithms, "*Proceedings 24th Asilomar Conference Signals, Systems, Computers,* Pacific Grove, CA., pp. 263-267, November 1990.

Widrow, B. et al.: "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter, " *Proceedings of the IEEE*, Vol. 64, pp. 1151-1162, August 1976.

Zirngibl, M., Joyner, C. H. et al.: "LARnet a Local Access Router Network," *IEEE Photonics Technology Letters*, Vol. 7, No. 2, pp. 215-217, February 1995.

VITA

Michael J.Minardi was born in Los Angeles, in March 1957 and grew up in Dayton Ohio. He received a B.E.E. and a B.S. in Mathematics (both summa cum laude and with deparmental honors) from the University of Dayton in 1979 and an M.S.E.E. from Georgia Institute of Technology in 1981.

He worked at the Georgia Tech Research Institute as a research engineer in the area of electronic countermeasures from 1979-1983 and 1989-1993. From 1983-1988 he worked for GTE Government Systems Division as a radar systems engineer at the ALTAIR deep space tracking radar located at the Kwajalein Missile Range in the Marshall Islands. In 1993 he went to work for the Radar Branch of the USAF Wright Laboratories, at which time he entered the Ph.D. program full time at Georgia Tech under the USAF "Palace Knight" program. He is a member of Tau Beta Pi and Pi Mu Epsilon.